Statistics for Biology and Health

Jianguo Sun

The Statistical Analysis of Interval-censored Failure Time Data

Springer

Jianguo Sun
Department of Statistics
University of Missouri
Columbia, MO 65211
USA
sunj@missouri.edu

Library of Congress Control Number: 2006922771

ISBN-10: 0-387-32905-6
ISBN-13: 978-0387-32905-5

Printed on acid-free paper.

Printed in the United States of America. (MP/POW)

9 8 7 6 5 4 3 2 1

springeronline.com

To Xianghuan, Ryan, and Nicholas

Preface

Interval censoring is a type of censoring that has become increasingly common in the areas that produce failure time data. In the past 20 years or so, a voluminous literature on the statistical analysis of interval-censored failure time data has appeared. The main purpose of this book is to collect and unify some statistical models and methods that have been proposed for analyzing failure time data in the presence of interval censoring.

A number of books have been written that provide excellent and comprehensive coverage of the statistical analysis of failure time data in the presence of right censoring. These include Cox and Oakes (1984), Fleming and Harrington (1991), Andersen et al. (1993), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), and Lawless (2003). In general, right-censored failure time data can be treated as a special case of interval-censored data, and some of the inference approaches for right-censored data can be directly, or with minor modifications, applied to the analysis of interval-censored data. However, most of the inference approaches for right-censored data are not appropriate for interval-censored data due to the fundamental differences between right censoring and interval censoring. The censoring mechanism behind interval censoring is much more complicated than that behind right censoring. For right-censored failure time data, substantial advances in the theory and development of modern statistical methods are due to the theory of counting processes. Because of the complexity and special structure of interval censoring, the same theory is not applicable to interval-censored data. The goal of this book is to complement the literature on right-censored data by presenting statistical models and methods specifically developed for interval-censored failure time data.

This book is intended to provide an up-to-date reference for those who are conducting research on the analysis of interval-censored failure time data as well as those who need to analyze interval-censored data to answer substantive questions. It can also be used as a text for a graduate course in statistics or biostatistics that has basic knowledge of probability and statistics as a prerequisite. The main focus is on methodology, and applications of the methods that are based on real data are given along with numerical calculations.

To keep the book at a reasonable length, some topics are discussed only briefly at the end of each chapter in the Bibliography, Discussion, and Remarks section or in the last chapter. Also, although some asymptotic results are discussed, their technical derivations are not presented. Because the literature on interval-censored data is extensive, the choice of subject matter is difficult. The material discussed in detail is to some extent a reflection of the author's interests in this field. However, our attempt has been to present a relatively complete and comprehensive coverage of the fundamental concepts along with selected topics in the field.

Chapter 1 contains introductory material and surveys basic concepts and regression models for the analysis of failure time data. Examples of right- and interval-censored survival data are discussed, and several types of interval censoring commonly seen in practice are described. Before considering the nonparametric and semiparametric approaches, which are the focus of the book, some parametric models and methods are presented in Chapter 2. Also, in Chapter 2, some imputation approaches are briefly investigated for the analysis of interval-censored failure time data.

Chapters 3 to 10 concern nonparametric and semiparametric approaches for interval-censored data. Chapter 3 considers statistical procedures for nonparametric estimation of survival and hazard functions, and Chapter 4 deals with nonparametric comparisons of survival functions. Both rank-based and survival-based procedures are investigated. Regression analysis of current status data, or case I interval-censored data, is discussed in Chapter 5, and Chapter 6 considers regression analysis of general, or case II interval-censored failure time data. The analysis of bivariate interval-censored failure time data is the subject of Chapter 7, which considers both nonparametric and semiparametric approaches. Chapter 8 deals with doubly censored failure time data. In this situation, the survival time of interest is the duration between two related events and the observations on the occurrences of both events could be right- or interval-censored. The analysis of event history data in the presence of interval censoring, which are commonly referred to as panel count data, is considered in Chapter 9. Chapter 10 contains brief discussions of several other important topics in the field for which it is not feasible to give a detailed discussion. These include regression diagnostics, regression analysis with interval-censored covariates, Bayesian inference approaches, and informative interval censoring.

In all chapters except Chapter 10, we have used references sparsely except in the last section of each chapter, which provides bibliographical notes including related references.

Many persons have contributed directly and indirectly to this book. First, I want to thank Diane Finkelstein, Jian Huang, Linxiong Li, Liuquan Sun, Tim Wright, and Ying Zhang for their many critical comments and suggestions. I am especially indebted to Tim Wright, who patiently read all the chapters and made numerous corrections in an early draft of the book. I owe my thanks to Do-Hwan Park, Xingwei Tong, Lianming Wang, Zhigang Zhang, Qiang Zhao, and Chao Zhu, who not only read parts of the draft and gave their important comments but also provided great computational help. Also, I would like to express my thanks to Nancy Flourney, our department chair, for her encouragement and support during this period, and Jack Kalbfleisch, Steve Lagakos, Jerry Lawless, and LJ Wei for their important influence on my academic life and their guidance in the early years of my research.

Finally, I thank my family and especially my wife, Xianghuan, for her patience and support during this project.

January 2006 *Jianguo Sun*

Contents

1

Introduction

1.1 Failure Time Data

By failure time data, we mean data that concern positive random variables representing times to certain events. Examples of the event, often referred to as the failure or survival event, include death, the onset of a disease or certain milestone, the failure of a mechanical component of a machine, or learning something. The occurrence of the event is usually referred to as a *failure*. Sometimes we also use the terminology survival data and refer to the variable of interest as survival time or the survival variable. Failure time data arise extensively in medical studies, but there are many other investigations that also produce failure time data. These include biological studies, demographical studies, economic and financial studies, epidemiological studies, psychological experiments, reliability experiments, and sociological studies.

The analysis of failure time data usually means addressing one of three problems. They are estimation of survival functions, comparison of treatments or survival functions, and assessment of covariate effects or the dependence of failure time on explanatory variables. We consider methods that can be used to deal with these problems for interval-censored data. A survival function, which is formally defined below, gives the probability that failure time is greater than a certain time and is of considerable interest in failure time analysis.

For a number of reasons, special methods are required to treat failure time data. One reason, which also is a major feature that distinguishes the analysis of failure time data from other statistical fields, is the existence of censoring, such as right censoring, which is discussed below. Censoring mechanisms can be quite complicated and thus necessitate special methods of treatment. The methods available for other types of data are usually simply not appropriate for censored data. Truncation is another feature of some failure time data that requires special treatments. We focus mainly on censoring and discuss only some special types of truncation. Before discussing censoring and truncation

Table 1.1. Remission times in weeks for acute leukemia patients

Group	Survival times in weeks
6-MP	6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^* 32^*, 32^*, 34^*, 35^*
Placebo	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

in more detail, we describe two examples to introduce failure time data and their features.

1.1.1 Remission Times of Acute Leukemia Patients

Table 1.1, reproduced from Freireich et al. (1963) and Gehan (1965), presents a typical set of failure time data arising from a clinical trial on acute leukemia patients. In the table, remission times in weeks are given for 42 patients in two treatment groups. One treatment is the drug 6-mercaptopurine (6-MP) and the other is the placebo treatment. The study was performed over a one-year period and the patients were enrolled into the study at different times. A primary concern is the comparison of the two treatments with respect to ability to maintain remission. In other words, it is of interest to know if the patients with drug 6-MP had significantly longer remission times than those given the placebo treatment.

For the observed information given in Table 1.1, the starred numbers are censoring times or censored remission times. That is, such an observation is the amount of time from when the patient entered the study to the end of the study. These remission times were censored because these patients were still in the state of remission at the end of the trial and thus their remission times were known only to be greater than the censoring times. For the other patients, their remission times were observed exactly. This situation commonly occurs in failure time studies, and the resulting data are usually referred to as right-censored failure time data. Note that for the comparison of the two treatments, a simple t-test is not applicable because it cannot handle the censored remission times, and certainly discarding these times is not desirable. For more discussion and the analysis of this data set, readers are referred to Kalbfleisch and Prentice (2002) in addition to Freireich et al. (1963) and Gehan (1965).

1.1.2 Times to the First Use of Marijuana

Turnbull and Weiss (1978) discussed a set of failure time data from a study on the use of marijuana by high school students, and the data are given in Table 1.2. In the study, 191 California high school boys were asked the question, "when did you first use marijuana?" As expected, some boys remembered the exact age when they first used it, and some boys used it but could not

Table 1.2. Ages in years to the first use of marijuana

Age	No. of exact observations	No. of left-censored observations	No. of right-censored observations
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2
15	13	18	3
16	3	14	2
17	1	6	3
18	0	0	1
>18	4	0	0

remember when they first used marijuana. Also there were boys who never used it.

Corresponding with these three situations, there are three types of observations about the age when marijuana was first used. For the first situation, the age is known exactly. For the second and third situations, the age is known only to be smaller or greater than the current age of the boy, and these types of observations are usually referred to as left-censored or right-censored observations, respectively. For the data set, one question of interest is to estimate the probability of having used marijuana at certain ages for high school boys. It is apparent that the simple empirical estimate is not appropriate unless one disregards some of the left- and right-censored observations. Among others, Klein and Moeschberger (2003) and Turnbull and Weiss (1978) analyzed this data set.

1.1.3 Censoring and Truncation

As mentioned above, censoring is one of the unique features of failure time data. By censoring, we mean that an observation on a survival time of interest is incomplete, that is, the survival time is observed only to fall into a certain range instead of being known exactly. Note that censored data are different from missing data as censored observations still provide some partial information, whereas missing observations provide no information about the variable of interest. Different types of censoring arise in practice, but the one that receives most of the attention in the literature is right censoring.

By right censoring or right-censored failure time data, we mean that the failure time of interest is observed either exactly or to be greater than a censoring time. A typical situation that yields right-censored observations is one in which a survival study has to end due to, for example, time constraints or resource limitations. In this case, for subjects whose survival events have not occurred at the end of the study, their survival times are not observed exactly but are known to be greater than the study end time, i.e., they are right-censored. For subjects who have already failed by the end of the study, their failure times are known exactly. Of course, the study end time could be different for different subjects, and some subjects may withdraw from the study before the end for some reasons. In a more general setting, which is appropriate in many applications, for each subject, there exists a censoring variable representing the right censoring time. If the survival variable is smaller than the censoring variable, the observation is exact and otherwise, it is right-censored. This is usually referred to as the random censorship model.

It is apparent that in general, one has to understand the way that right censoring occurs to analyze right-censored failure time data properly. To simplify the analysis, an independent right censoring mechanism is commonly assumed. By this, we mean that the failure rate or hazard is the same for the subjects who are still in the study and the subjects who have been censored out. More specifically, under independent right censoring, we have that

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t, Y(t) = 1)}{\Delta t}$$

(Kalbfleisch and Prentice, 2002), where T denotes the survival variable of interest, and $Y(t) = 1$ means that the subject has neither failed nor been censored prior to time t. Under the random censorship model, the above condition is equivalent to

$$\lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \le T < t + \Delta t | T \ge t, C \ge t)}{\Delta t} ,$$

where C denotes the censoring variable.

There exist different types of right censoring as well as other types of censoring. For example, the censoring mechanism that stops the study at the same fixed time point for all subjects is usually referred to as Type 1 censoring. Type 2 censoring means that the study stops if a prespecified number of individuals out of all study individuals have failed. In addition to right censoring, some observations may be left-censored, meaning that the failure time is known only to be less than certain time. Interval censoring, the focus of this book, is introduced in the next section.

Truncation refers to situations where a subject is included in a study only if the corresponding failure time satisfies certain conditions. A simple and common example that yields truncated failure time data is a cohort study in which subjects are included in the study only if they experience some initial event prior to the survival event. In this case, for all subjects in the study, their failure times are greater than the occurrence times of the initial event. This type of truncation is commonly referred to as left-truncation. Independent truncation can be defined similarly to independent right censoring and is usually assumed for the analysis of truncated failure time data. For a more detailed discussion of right censoring and truncation, among others, see Kalbfleisch and Prentice (2002) and Lawless (2003). They give various statistical methods for the analysis of right-censored failure time data such as those discussed in Section 1.1.1.

1.2 Failure Time Data with Interval Censoring

As discussed in the previous section, failure time data occur in many ways and in many fields, and there are a number of reasons why special methods are needed for their analyses. One key feature of failure time data is censoring, and there exist many excellent books on right censoring. Here we focus on interval censoring, which is more challenging than right censoring, and for such data the methods developed for right censoring do not generally apply.

By interval censoring, we mean that study subjects or failure time processes of interest are not under continuous observation. As a consequence, the failure or survival time is not always exactly observed or right-censored. For an interval-censored observation, one only knows a window, that is, an interval, within which the survival event has occurred. Exact or right-censored failure times can be regarded a special case of interval-censored failure times as in such cases, the interval reduces to a single point or is unbounded on the right. More generally, one could define an interval-censored observation as a union of several nonoverlapping windows or intervals (Turnbull, 1976).

Interval-censored failure time data occur in many areas including demographical, epidemiological, financial, medical, and sociological studies. A typical example of interval-censored data occurs in medical or health studies that entail periodic follow-ups, and many clinical trials and longitudinal studies fall into this category. In such situations, interval-censored data may arise in several ways. For instance, an individual may miss one or more observation times that have been scheduled to clinically observe possible changes in disease status and then return with a changed status. Alternatively, individuals may visit clinical centers at times that are convenient to them rather than at predetermined observation times. In both situations, the data on change in status are interval-censored. Even if all study subjects follow exactly the predetermined observation schedule, one still cannot observe the exact time of the occurrence of the change of the status assuming that it is a continuous variable. In the last situation, one has grouped failure time data, that is, interval-censored data for which the observation for each subject is a member of a collection of nonoverlapping intervals. Grouped failure time data can be dealt with relatively easily. Among others, Lawless (2003) discussed this type of failure time data. In the following, we focus on interval-censored data that are not grouped failure time data.

We present several examples below to further illustrate some of the general concepts, definitions, common features, and the structure of interval-censored data. The first two examples concern univariate failure time variables representing the time from the beginning of a study to the occurrence of an event of

Table 1.3. Death times in days for 144 male RFM mice with lung tumors

Group	Tumor status	Death times
CE	With tumor	381, 477, 485, 515, 539, 563, 565, 582, 603, 616, 624, 650 651, 656, 659, 672, 679, 698, 702, 709, 723, 731, 775, 779 795, 811, 839
	No tumor	45, 198, 215, 217, 257, 262, 266, 371, 431, 447, 454, 459 475, 479, 484, 500, 502, 503, 505, 508, 516, 531, 541, 553 556, 570, 572, 575, 577, 585, 588, 594, 600, 601, 608, 614 616, 632, 632, 638, 642, 642, 642, 644, 644, 647, 647, 653 659, 660, 662, 663, 667, 667, 673, 673, 677, 689, 693, 718 720, 721, 728, 760, 762, 773, 777, 815, 886
GE	With tumor	546, 609, 692, 692, 710, 752, 773, 781, 782, 789, 808, 810 814, 842, 846, 851, 871, 873, 876, 888, 888, 890, 894, 896 911, 913, 914, 914, 916, 921, 921, 926, 936, 945, 1008
	No tumor	412, 524, 647, 648, 695, 785, 814, 817, 851, 880, 913, 942 986

interest. The third example is about a univariate failure time variable representing the duration between two related events. The fourth example contains two correlated failure times of interest.

1.2.1 Lung Tumor Data

Hoel and Walberg (1972) give a set of data for 144 male RFM mice in a tumorigenicity experiment that involves lung tumors. The data are presented in Table 1.3 and consist of the death time of each animal measured in days and an indicator of lung tumor presence (1) or absence (0) at time of death. The experiment involves two treatments, conventional environment (CE, 96 mice) and germ-free environment (GE, 48 mice). Lung tumors in RFM mice are predominantly nonlethal, meaning that the occurrence of a tumor does not change the death rate.

Tumorigenicity experiments are usually designed to determine whether a suspected agent or environment accelerates the time until tumor onset in experimental animals. In these situations, the time to tumor onset is usually of interest but not directly observable. Instead, only the death or sacrifice time of an animal is observed, and the presence or absence of a tumor at the time is known. If the tumor can be considered to be rapidly lethal, meaning that its occurrence kills the animal right away, it is reasonable to treat the time of death or sacrifice of an animal as an exact or right-censored observation of the tumor onset time. In this case, the data can be analyzed by methods developed for right-censored failure time data. On the other hand, if the tumor is nonlethal as that considered here, then the time to tumor onset is only known to be less than or greater than the observed time of death or sacrifice. In other words, only left- or right-censored observations on the tumor onset time are available, and the tumor onset time is interval-censored. This type of interval-censored data is commonly referred to as current status data (see Section 1.3.1).

Among others, one common objective of tumorigenicity experiments is to investigate the effect of a suspected agent or environment on tumor prevalences or incidence rates. For the data in Table 1.3, for example, it is of interest to compare the lung tumor incidence rates of the two treatment groups. More discussion and the analysis of this data set are given in Sections 3.2, 4.5.1, 5.2.2, 5.3.2, 5.4.2, and 5.5.2.

1.2.2 Breast Cancer Study

Table 1.4 presents data from a retrospective study on early breast cancer patients who had been treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. The data are reproduced from Finkelstein and Wolfe (1985) and consist of 94 patients who were given either radiation therapy alone (RT, 46) or radiation therapy plus adjuvant chemotherapy (RCT, 48).

In the study, patients were supposed to be seen at clinic visits every 4 to 6 months. However, actual visit times differ from patient to patient, and times between visits also vary. At visits, physicians evaluated the cosmetic appearance of the patient such as breast retraction, a response that has a negative impact on overall cosmetic appearance. The goal of the study is to compare the two treatments, radiation therapy alone and radiation therapy plus adjuvant chemotherapy, with respect to their cosmetic effects. Adjuvant chemotherapy improves the relapse-free and overall survival for some patients. But there exists some experimental and clinical evidence that suggests that chemotherapy intensifies the acute response of normal tissue to radiation treatment.

The data contain information about the time to breast retraction. However, no exact time was observed. There are 38 patients who did not experience breast retraction during the study, giving right-censored observations

Table 1.4. Observed intervals in months for times to breast retraction of early breast cancer patients

Group	Observed intervals in months
RT	(45,], (25,37], (37,], (4,11], (17,25], (6,10], (46,], (0,5], (33,], (15,], (0,7], (26,40], (18,], (46,], (19,26], (46,], (46,], (24,], (11,15], (11,18] (46,], (27,34], (36,], (37,], (22,], (7,16], (36,44], (5,12], (38,], (34,] (17,], (46,], (19,35], (46,], (5,12], (9,14], (36,48], (17,25], (36,], (46,] (37,44], (37,], (24,], (0,8], (40,], (33,]
RCT	(8,12], (0,5], (30,34], (16,20], (13,], (0,22], (5,8], (13,], (30,36], (18,25] (24,31], (12,20], (10,17], (17,24], (18,24], (17,27], (11,], (8,21], (17,26], (35,] (17,23], (33,40], (4,9], (16,60], (33,], (24,30], (31,], (11,], (15,22], (35,39] (16,24], (13,39], (15,19], (23,], (11,17], (13,], (19,32], (4,8], (22,], (44,48] (11,13], (34,], (34,], (22,32], (11,20], (14,17], (10,35], (48,]

denoted by the intervals with no right end points. For the other patients, the observations are intervals, representing the time periods during which breast retractions occurred. The intervals are given by the last clinic visit time at which breast retraction had not yet occurred and the first clinic visit time at which breast retraction was detected. For example, the observation $(6, 10]$ means that at month 6, the patient had shown no deterioration in cosmetic state, but by the next visit at month 10, breast retraction was present. That is, we have interval-censored data for the time to breast retraction. The analysis of this data set is discussed in Sections 2.3.4, 2.4.3, 3.4.4, 4.5.2, 6.2.3, and 6.5.3.

1.2.3 AIDS Cohort Study

Table 1.5 gives a set of data arising from a cohort study of 257 individuals with Type A or B hemophilia and is reproduced from Kim et al. (1993). The subjects in the study were treated at two French hospitals beginning in 1978 and were at risk for infection of the human immunodeficiency virus (HIV) through contaminated blood factor received for their treatments. The table includes only 188 subjects who were found to be infected with HIV during the study period that lasted from 1978 to August 1988. Among these infected patients, 41 subsequently progressed during the study to the acquired immunodeficiency syndrome (AIDS) or related clinical symptoms, which will be simply referred to as an AIDS diagnosis. One variable of great interest in this study, and also in other similar studies, is the time from HIV infection (or more precisely HIV seroconversion) to AIDS diagnosis. It is often referred to as AIDS incubation or latency time. The AIDS latency time provides information about HIV infection progression and plays an important role in, for example, predicting HIV prevalences.

In this study of HIV infection times, only intervals that bracket the infection time for each study subject are available. This is because HIV infection status was determined by retrospective tests of stored blood sera, and thus the exact HIV infection time was not observed. The intervals given in Table 1.5 are formed by the times at which the last negative and first positive test results were obtained with a unit of six months. In terms of AIDS diagnosis times, they either were observed exactly (for 41 subjects with AIDS diagnosis before the collection of the data) or were right-censored (for the other subjects). This type of censored data is usually referred to as doubly censored failure time data. Note that in the original data set, there are a few subjects whose AIDS diagnosis times were given by narrow intervals, and these are not included in Table 1.5 for simplicity.

In addition to HIV infection and AIDS diagnosis times, Table 1.5 also includes information on a covariate that is a group indicator. The subjects in the study were classified into two groups according to the amount of blood factor that they received. The heavily treated group includes the individuals who received at least 1000 μg/kg of the blood factor for at least one year

Table 1.5. Observed intervals in 6-month scale given by $(L, R]$ for HIV infection time and observations (denoted by T with starred numbers being right-censored times) for AIDS diagnosis time for 188 HIV-infected patients (the numbers in parentheses are multiplicities)

L	R	T		L	R	T		L	R	T		L	R	T		L	R	T	
Lightly treated group																			
0	5	23*	(2)	0	11	23*	(2)	0	12	23*	(3)	0	14	23*		11	14	23*	(3)
0	15	23*	(9)	0	16	23*	(4)	0	17	23*		0	18	23*		11	15	23*	
2	10	23*		5	8	23*		6	10	23*		6	12	23*		13	16	23*	
7	12	23*		7	13	23*		7	15	23*		8	13	23*		5	13	21	
8	14	23*	(3)	9	12	23*	(2)	9	16	23*		10	14	23*	(4)	10	12	23*	(2)
11	13	23*	(4)	11	14	23*		12	14	23*	(4)	12	15	23*	(3)	10	15	23*	
13	15	23*	(4)	14	16	23*	(5)	0	3	8		0	12	15		10	16	23*	
5	12	16		9	11	20		9	12	21		10	12	20		2	16	21	
12	13	22		12	15	22		0	13	23*		6	13	17		12	14	20	
3	11	23*		4	11	23*		5	13	23*		7	16	23*		7	16	21	
8	12	23*		9	15	23*		11	13	23									
Heavily treated group																			
0	7	23*		0	11	23*		0	12	23*	(2)	0	13	23*		0	7	16	
0	14	23*	(3)	0	15	23*	(2)	0	16	23*		2	14	23*		8	11	18	
4	7	23*	(2)	6	9	23*		6	10	23*		7	10	23*		9	12	16	
8	10	23*	(2)	8	12	23*	(3)	9	11	23*	(7)	9	12	23*	(2)	9	14	16	
9	15	23*		10	12	23*		10	13	23*	(4)	11	13	23*	(7)	7	15	23*	
11	14	23*	(2)	12	15	23*	(3)	12	16	23*		13	15	23*	(8)	0	13	23*	
13	16	23*	(2)	14	16	23*	(5)	0	7	13		0	10	12		2	15	23*	
0	15	21		2	7	17		4	7	12		4	8	13		6	15	23*	
6	9	19		7	10	15		8	12	18		8	12	22		12	14	18	
8	13	15		8	13	18		9	11	15		9	11	16		12	15	18	(2)
9	12	17		9	12	23		11	13	20		12	14	20		12	14	21	
13	15	23																	

between 1982 and 1985, whereas the subjects in the lightly treated group received less than 1000 μg/kg in each year. Among others, one objective of interest in this type of study is to estimate the distribution of AIDS latency time. One could also be interested in investigating the effect of covariates on the distribution of the AIDS latency time. These are discussed in detail in Section 8.5.2.

1.2.4 AIDS Clinical Trial

Goggins and Finkelstein (2000) discussed a data set arising from an AIDS clinical trial, AIDS Clinical Trial Group (ACTG) 181, on HIV-infected individuals. The study is a natural history substudy of a comparative clinical trial of three anti-pneumocystis drugs and concerns the opportunistic infection cytomegalovirus (CMV). During the study, among other activities, blood and urine samples were collected from the patients at their clinical visits and tested for the presence of CMV, which is also commonly referred to as shedding of the virus. These samples and tests provide observed information on the two variables, the times to CMV shedding in blood and in urine.

The observed information is presented in data set I of Appendix A and contains the observed intervals for the times to CMV shedding in blood and urine from 204 patients who provided at least one urine and blood samples during the study. Some intervals contain time zero, that is, the shedding times are left-censored because the shedding had already occurred for these patients when they entered the study. Some intervals have no right end points, that is, the shedding times are right-censored because the corresponding patients had not yet started shedding by the end of the study. For the other patients, their observed intervals are given by the last negative and first positive blood and urine tests, respectively. In summary, we have two possibly correlated failure times of interest, and observations on both of them are interval-censored.

In addition to the observed information about CMV shedding times in blood and in urine, the data set also includes information about the patient's baseline CD4 cell counts given by the indicator variable CD4.ind. In particular, the patients are classified into two groups with CD4.ind equal to 1 if the baseline CD4 cell count was less than 75 (cells/μl) and 0 otherwise. The CD4 cell count indicates the status of a person's immune system and is commonly used to measure the stage of HIV infection. For this data set, one problem of interest is to estimate the association between CMV shedding times in blood and in urine or the joint distribution of the times to CMV shedding in blood and in urine. It is also often of considerable interest to determine the relationship between the time to CMV shedding and the baseline CD4 cell count or whether the baseline CD4 cell count is predictive of CMV shedding in either blood or urine. The analysis of this data set is discussed in Sections 7.2.3 and 7.4.3l.

More examples of interval-censored failure time data and their analyses are given throughout the book. In the next section, we formally introduce several types of interval-censored data that are commonly seen in practice and their corresponding formulations. The methods for their analyses are discussed in the following chapters.

1.3 Types of Interval Censoring and Their Formulations

Let T be a nonnegative random variable representing the failure time of an individual in a failure time study. An observation on T is interval-censored if instead of observing T exactly, only an interval $(L\,,\,R]$ is observed such that

$$T \in (\,L\,,\,R\,]\,, \tag{1.1}$$

where $L \leq R$. In the following, we use the convention that $L = R$ means an exact observation, and $R = \infty$ represents a right-censored observation.

In this book, four types of interval censoring that commonly occur in practice and their analyses are considered in detail.

1.3.1 Case I Interval-censored Failure Time Data

The term *case I interval-censored data* is commonly used to refer to interval-censored failure time data in which all observed intervals "include" either time zero or infinity (Groeneboom and Wellner, 1992; Huang, 1996). In other words, the observation on each individual failure time is either left- or right-censored, that is, either $L = 0$ or $R = \infty$. Case I interval-censored data occur when each study subject is observed only once and the only observed information for the survival event of interest is whether the event has occurred no later than the observation time. Instead of the intervals in (1.1), a more convenient representation of case I interval-censored data is $\{ C , \delta = I(T \leq C) \}$, where C denotes the observation time and I is the indicator function. Note that case I interval-censored data differ from right-censored data or left-censored data, which usually include some failure times that are observed exactly.

Case I interval-censored data are also often referred to as *current status data*, a term originating from demographical studies. Cross-sectional studies and tumorigenicity experiments on nonlethal tumors are two types of studies that frequently produce case I interval-censored data. The former is commonly used in demographical studies, and the lung tumor study discussed in Section 1.2.1 provides an example of the latter. Note that there is a fundamental difference between the current status data arising from these two types of studies although they are analyzed in the same way. The current status data from the former occur mainly due to study designs, whereas those given in the latter are observed usually due to the inability to measure the variable directly and/or accurately.

1.3.2 Case II Interval-censored Failure Time Data

Interval-censored data that include at least one interval $(L , R]$ with both L and R belonging to $(0, \infty)$ are usually referred to as *general or case II interval-censored data* (Groeneboom and Wellner, 1992; Huang and Wellner, 1997; Sun, 1998, 2005). In other words, case II interval-censored data are interval-censored data that include some finite intervals away from zero.

Another way to represent a case II interval-censored observation is to use

$$\{ U , V , \delta_1 = I(T \leq U) , \delta_2 = I(U < T \leq V) , \delta_3 = 1 - \delta_1 - \delta_2 \} \quad (1.2)$$

assuming that each subject is observed twice, where U and V are two random variables satisfying $U \leq V$ with probability 1. This formulation is convenient and often used, for example, in a theoretical investigation of an inference procedure. Note that by taking $U = V = C$, case I interval-censored data can be described by (1.2). Yu et al. (2000) generalize this formulation to include exact observations. Note that in the literature, the term case II interval-censored data is sometimes used to refer only interval-censored data that are given in representation (1.2).

Another generalization of the formulation (1.2) is to assume that there exists a set of observation time points, say $U_1 \leq U_2 ... \leq U_K$, for each study subject, where K is a random integer. The observed information then has the form

$$\{ (K , U_j , \delta_j = I(U_{j-1} < T \leq U_j)) , j = 1, ..., K \} , \tag{1.3}$$

where $U_0 = 0$. This formulation or type of failure time data is often referred to as case K or mixed case interval-censored data (Schick and Yu, 2000; Wellner, 1995). It is apparent that the above formulation includes the representation (1.2) as a special case and provides a natural representation of interval-censored failure time data arising from longitudinal studies with periodic follow-up.

All three representations, (1.1) to (1.3), give rise to the same likelihood function. Note that although both representations (1.2) and (1.3) seem natural, it is not common to have interval-censored data collected or given in these formats in practice. However, it is much easier and more natural to impose assumptions such as independence with T on them than on representation (1.1), which is often needed for derivation of the asymptotic properties of inference procedures. For data given in representation (1.2) or (1.3), one can easily obtain the corresponding data with representation (1.1). On the other hand, it is apparently impossible to transform representation (1.1) to (1.3) without extra information about observation process, and it is not straightforward to transform observations given in representation (1.1) to these in the representation (1.2). More discussion on this is given later. In the following chapters, we mainly focus on the first two representations and use them interchangeably.

1.3.3 Doubly Censored Failure Time Data

Consider a survival study involving two related events and let X and S denote the times of the occurrences of the two events with $X \leq S$. Define $T = S - X$ and suppose that T is the survival time of interest. By doubly censored failure time data, we mean that the observations on both X and S are interval-censored (De Gruttola and Lagakos, 1989; Sun, 2004). Specifically, suppose that instead of observing X and S exactly, one only observes two intervals $(L , R]$ and $(U , V]$ such that

$$X \in (L , R] \ , \ S \in (U , V] ,$$

where $L \leq R$ and $U \leq V$ with probability 1. In other words, the observations on T are doubly censored.

The special type of doubly censored data in which S is only right-censored occurs commonly, and in this case, one has either $U = V$ or $V = \infty$. Another formulation for this special case that may be more natural is to assume that there exists a censoring variable C, which is often assumed to be independent of S. The observation on S then consists of $S^* = \min\{S, C\}$ and $\delta = I(S^* = S)$, where I is the indicator function as before.

One often sees doubly censored failure time data in disease progression studies where the two events may represent infection and subsequent onset of a certain disease, respectively, such as the example discussed in Section 1.2.3. In these situations, doubly censored observations occur mainly due to the nature of the disease and/or the structure of the study design. In the example given in Section 1.2.3, X and S represent HIV infection and AIDS diagnosis times, respectively, and T is the AIDS latency time. For most AIDS cohort studies, as in this example, because HIV infection usually is determined through periodic blood tests, observations on it are commonly interval-censored. Also, observations on the diagnosis of AIDS could be, for example, right-censored due to the end of the study, thus yielding doubly censored data on T.

Doubly censored failure time data include as special cases right-censored and interval-censored failure time data. For example, they reduce to interval-censored data if the time of occurrence of the first event, X, can be observed exactly ($L = R$). Furthermore, if the observation on the time of occurrence of the subsequent event, S, is exact or right-censored, we then have a right-censored observation on T. Note that for doubly censored data, if X is observed exactly, for inferences about T, one may relabel so that $X = 0$, which typically is done in failure time data analysis

In the literature, doubly censored data considered here are sometimes referred to as doubly interval-censored data (Sun, 1995) to distinguish them from another type of doubly censored failure time data. In the latter, the survival time of interest is observed exactly if it is within a window and left- or right-censored if it is to the left or right of the window (Cai and Cheng, 2004; Chen and Zhou, 2003; Turnbull, 1974). A key difference between the two types of data is that for the latter type of data, some exact failure times are observed, but if not, they become case I interval-censored data. The methods required for the analyses of these two types of doubly censored data are different.

1.3.4 Panel Count Data

Interval censoring occurs in a more general setting than survival studies. In failure time data analysis, the random variable of interest is always the time to an event, and the event is treated as an absorbing event. In other words, the event can occur only once such as fatal failure or death. In practice, however, there exist many situations where the event of interest can occur multiple times such as a tumor or disease symptom. In these situations, in addition to the time to the event or between the occurrences of the event, one may also want to study the occurrence process of the event. Without interval censoring, that is, if the process is observed continuously, then one has what is commonly called *recurrent event data*, in which one knows all the exact occurrence times of the event (Cai and Schaubel, 2004; Chang and Wang, 1999). In the presence of interval censoring, which arises if the subject or occurrence process is observed only at discrete time points, one only knows the numbers of the occurrences of the event between observation times. In this case, the observed data are often referred to as *panel count data* (Kalbfleisch and Lawless, 1985; Sun and Wei, 2000). However, if the event can occur only once, then the data become interval-censored failure time data. Panel count data are also sometimes referred to as interval count data or interval-censored recurrent event data (Lawless and Zhan, 1998; Thall, 1988).

Panel count data frequently occur in long-term clinical, industrial, or animal studies. In a follow-up cancer study, for example, one could be interested in the recurrence rate of one or more types of tumors or of tumors at one or more locations. For such a study, it is usually impossible or impractical to follow study subjects continuously, and thus panel count data are obtained. Another example is longitudinal sociological studies on, for example, job changes.

Define a counting process $N(t)$ with $N(t)$ denoting the number of occurrences of a recurrent event up to and including time t. For usual survival problems, $N(t)$ is a 0-1 counting process, and the counting process formulation has been used extensively in the literature for the development of statistical methods for the analysis of right-censored failure time data. For more detailed discussion on this, one can read, for example, the book by Andersen et al. (1993). The methodology described there can also be used for the analysis of recurrent event data. In the case of panel count data, the values of $N(t)$ are known only at different observation time points, and we do not know the time points at which $N(t)$ jumps. In this book, for the analyses of panel count data, we focus on methods that allow observation times to vary from subject to subject.

1.3.5 Independent Interval Censoring, Notation, and Remarks

By *independent interval censoring*, as independent right censoring, we mean that the mechanism that generates the censoring is independent of the underlying variable of interest completely or given covariates. For current status data, this implies that C and T are independent. For interval-censored data given in representation (1.2) or (1.3), the independent interval censoring means that the joint distribution of U and V or the U_j's contains no parameters that are involved in the survival function of T. With respect to the data given in the format (1.1), the independent interval censoring assumes that an interval $(L, R]$ gives no more than the information that T is simply bracketed by the two observed values. In other words, we have

$$P(T \le t \,|\, L = l, R = r, L \le T < R) = P(T \le t \,|\, l \le T < r)$$

(Self and Grossman, 1986; Zhang et al., 2005), or

$$P(\,L < T \leq R\,|\,L = l, R = r\,) \;=\; P(\,l < T \leq r\,)$$

and the joint distribution of L and R is free of the parameters involved in the survival function of T. More remarks about this independent censoring mechanism are given in Section 10.5 along with discussion on situations where it does not hold. Under the independent interval censoring, one does not have to deal with the censoring mechanism in analyzing interval-censored data. Throughout the book, the independent interval censoring is assumed unless otherwise specified.

For presentation of an interval-censored observation, instead of $(\,L\,,\,R\,]$, one could also use $[\,L\,,\,R\,]$, $[\,L\,,\,R\,)$, or $(\,L\,,\,R\,)$ (Peto, 1973; Turnbull, 1976). If T is continuous, it is apparent that there is no difference among them in the sense that they represent the same observed information about T. On the other hand, if T is discrete, care is needed because the information given by them can be different. Some discussion on this can be found in Ng (2002), and the notation $(\,L\,,\,R\,]$ is used throughout this book.

As mentioned above, for T, exact and right-censored observations can be seen as special cases of interval-censored observations. In practice, a set of interval-censored data may include both exact and purely interval-censored observations. Suppose that T is continuous. Then for an exact observation $T = t_0$, its likelihood contribution is $f(t_0)$, and for a purely interval-censored observation $(\,L\,,\,R\,]$, the likelihood contribution has the form $S(L) - S(R)$, where $f(t)$ and $S(t) = P(T > t)$ denote the density and survival functions of T, respectively. In the following, we mainly focus attention on purely interval-censored observations and the corresponding likelihood contribution in the construction of likelihood functions. In other words, for the construction of likelihood functions, we assume for convenience that no exact observations are present unless otherwise specified. The derivation and development of most likelihood-based inference procedures in this book hold when exact failure times are present and the corresponding likelihood contributions are included.

In addition to those described in the previous subsections, interval censoring can also occur in other formulations. For example, interval-censored data can arise from a multi-state model (Commenges, 2003). Also in a survival study, the variable that suffers interval censoring may be a covariate instead of the survival time of interest as discussed above (Goggins et al., 1999b). More generally, observations on both covariates and survival variables may be interval-censored (Zhao et al., 2005). More discussion on this can be found in Section 10.3. As in the case of right censoring, truncation may occur together with interval censoring. By truncation, as before, we mean that a subject is included in a study only if its failure time belongs to a certain window. Here truncation can occur for the same reasons as those for right-censored failure time data. For example, left-truncated and interval-censored data occur if the survival time T is observed only if T is greater than a certain value and only an interval to which T belongs can be observed. In the following, we focus mainly on situations without truncation.

We remark that in practice, interval-censored data are often collected and presented as discrete data, and this is especially the case when the data arise from follow-up studies with day, month, or year as the time unit. Therefore, it is natural and convenient to treat the underlying survival variables as discrete variables in the development of approaches for their analyses. Also it is reasonable and sometimes convenient to treat them as continuous variables as the measured values are often approximations to the true values due to, for example, measurement errors. This is especially the case for the investigation of large sample properties of the methods of analysis. In the following discussion, the two formulations are used interchangeably depending on convenience and purpose.

1.4 Concepts and Some Regression Models

Let T denote a nonnegative random variable representing the failure time of a subject, that is, the survival variable of interest. For inferences about T, the survival function and the hazard function are particularly useful for modeling. The survival function of T is defined as the probability that T exceeds a value t. Let $S(t)$ denote the survival function of T. Then one has

$$S(t) \,=\, P(\,T \,>\, t\,)\,, \;\; 0 \,<\, t \,<\, \infty\,.$$

The hazard function is defined differently for continuous and discrete survival variables and these definitions are given below. The probability density and distribution functions are often used too in survival analysis although not as frequently as the survival and hazard functions.

In addition to reviewing these functions along with their relationships, this section describes several continuous semiparametric regression models commonly used in survival analysis. These include the Cox or proportional hazards model, the proportional odds model, the additive hazards model, the accelerated failure time model, and the linear transformation model. Two discrete regression models are also presented. Some commonly used parametric models are discussed in the next chapter along with the corresponding inference procedures and the imputation approach for the analysis of interval-censored data.

1.4.1 Continuous Survival Variables

Assume that T is absolutely continuous and thus its probability density function $f(t)$ exists. By definition, it is easy to see that the density function and the survival function satisfy

$$f(t) \,=\, -\,dS(t)/\,dt$$

or

$$S(t) \;=\; \int_t^\infty f(s)\,ds\;.$$

The hazard function of T at time t is defined as

$$\lambda(t) \;=\; \lim_{\Delta t \to 0^+} \frac{P(\,t \le T < t + \Delta t \,|\, T \ge t\,)}{\Delta t}\;.$$

It represents the instantaneous probability that a subject fails at time t given that the subject has not failed before t. The survival, density, and hazard functions have one-to-one relationship. Specifically, given the density or survival function, we have

$$\lambda(t) \;=\; \frac{f(t)}{S(t)} \;=\; -\,\frac{d \log S(t)}{dt}\;.$$

On the other hand, it can be proved that

$$S(t) \;=\; \exp\left[-\int_0^t \lambda(s)\,ds\right] \;=\; \exp\left[\,-\,\Lambda(t)\,\right]$$

and

$$f(t) \;=\; \lambda(t)\,\exp[\,-\,\Lambda(t)\,]\,,$$

where $\Lambda(t) \;=\; \int_0^t \lambda(s)\,ds$, which is commonly referred to as the cumulative hazard function of T.

1.4.2 Discrete Survival Variables

Assume that T is a discrete survival variable taking values $s_1 < s_2 < ...$ with probability function $\{\,f(s_j) = P(T = s_j)\,;\, j = 1, 2, ...\}$. Then one has

$$S(t) \;=\; \sum_{j : t < s_j} f(s_j)\;.$$

In this case, the hazard of T at s_j is defined as

$$p_j \;=\; P(\,T = s_j \,|\, T \ge s_j\,) \;=\; \frac{f(s_j)}{S(s_j -)}\,,$$

the conditional probability that the failure occurs at s_j given that the failure has not occurred before s_j, $j \;=\; 1, 2, ...$

As in the continuous case, the survival, density, and hazard functions uniquely determine each other. Based on the above definitions, one can show that

$$S(t) \;=\; \prod_{j : t \ge s_j} (1 \,-\, p_j)$$

and

$$f(s_j) = p_j \prod_{l=1}^{j-1} (1 - p_l) .$$

In survival analysis, due to the special structure of the observed information and questions of interest, it is more convenient to model the hazard function or the survival function than other functions that determine the distribution of T. The remainder of this section discusses several such regression models.

1.4.3 The Proportional Hazards Model

Let $\boldsymbol{Z}$ be a vector of covariates including, for example, treatment indicator, age, and gender. As remarked before, a regression analysis provides an assessment of covariate effects on failure time, which is one of the important tasks in survival analysis. For this, a regression model is usually needed to specify how the covariates affect the failure time of interest. The proportional hazards (PH) or Cox model assumes that the hazard function of T has the form

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) \exp(\boldsymbol{Z}' \boldsymbol{\beta}) \tag{1.4}$$

given covariates $\boldsymbol{Z}$ (Cox, 1972). In the above, $\lambda_0(t)$ is an arbitrary unspecified baseline hazard function, and $\boldsymbol{\beta}$ is the vector of regression parameters. This model specifies that the covariates act multiplicatively on the hazard function.

The model (1.4) says that the ratio of the hazard functions for two subjects with different covariates is constant. In particular, for the two-sample situation where $Z = 0$ or 1, one has

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} = \exp(\beta) .$$

Under the PH model, the conditional density and survival functions of T given $\boldsymbol{Z}$ have the forms

$$f(t; \boldsymbol{Z}) = \lambda_0(t) \exp(\boldsymbol{Z}' \boldsymbol{\beta}) \exp\left[- \Lambda_0(t) \exp(\boldsymbol{Z}' \boldsymbol{\beta}) \right]$$

and

$$S(t; \boldsymbol{Z}) = \exp[-\Lambda_0(t) \exp(\boldsymbol{Z}' \boldsymbol{\beta})] = [S_0(t)]^{\exp(\boldsymbol{Z}' \boldsymbol{\beta})} ,$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(s) ds$$

and

$$S_0(t) = \exp\left[- \int_0^t \lambda_0(s) ds \right]$$

are the baseline cumulative hazard function and the baseline survival function. The conditional cumulative hazard function of T given $\boldsymbol{Z}$ has the form

$$\Lambda(t; \boldsymbol{Z}) \,=\, \Lambda_0(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta})\,.$$

The PH model is perhaps the most commonly used regression model in failure time data analysis. One main reason is that a simple and efficient inference procedure, the partial likelihood approach, about the regression parameter $\boldsymbol{\beta}$ is available for right-censored failure time data. The partial likelihood approach was proposed by Cox (1972, 1975) and has been studied by many authors. It is simple partly because the partial likelihood function used for such inferences is only a function of $\boldsymbol{\beta}$, and thus one does not have to deal with the baseline hazard function $\lambda_0(t)$. The approach is efficient because the resulting estimator of $\boldsymbol{\beta}$ is asymptotically equivalent to the estimator of $\boldsymbol{\beta}$ given by the full likelihood function. In addition to Cox (1972, 1975), Cox and Oakes (1984) and Kalbfleisch and Prentice (2002) give other references that discuss model (1.4) and its use in regression analyses of right-censored failure time data.

Many generalizations of the PH model exist. One allows $\boldsymbol{Z}$ to depend on time, which could be the case if, for example, $\boldsymbol{Z}$ includes the level of air pollution or the amount that a person exercises. Another allows the baseline hazard function to be different for subjects from different subgroups or subpopulations. To be specific, suppose that the population is divided into k strata and the hazard function of T for a subject from the jth stratum has the form

$$\lambda(t; \boldsymbol{Z}) \,=\, \lambda_{0j}(t)\, \exp(\boldsymbol{Z}'\boldsymbol{\beta})$$

given covariates $\boldsymbol{Z}$, $j \,=\, 1, ..., k$. That is, the hazard function may have different shapes for subjects from different stratum. Other generalizations include combinations of the two generalizations above and the use of non-linear relationships for the covariate effect rather than the linear relationship in (1.4) (Huang, 1999a). This book focuses mainly on time-independent covariates and related approaches to inferences.

1.4.4 The Proportional Odds Model

The proportional odds model is another regression model commonly used in survival analysis. It models the conditional survival function given covariates $\boldsymbol{Z}$ by postulating that

$$\frac{S(t; \boldsymbol{Z})}{1 \,-\, S(t; \boldsymbol{Z})} \,=\, e^{-\,\boldsymbol{Z}'\boldsymbol{\beta}}\, \frac{S_0(t)}{1 \,-\, S_0(t)}\,, \tag{1.5}$$

or

$$\mathrm{logit}[S(t; \boldsymbol{Z})] \,=\, \mathrm{logit}[S_0(t)] \,-\, \boldsymbol{Z}'\boldsymbol{\beta}\,.$$

As before, $S_0(t)$ denotes the baseline survival function or the survival function for the subjects with $\boldsymbol{Z} \,=\, 0$, and $\mathrm{logit}(x) \,=\, \log(x/(1-x))$.

As with the PH model, (1.5) also assumes that the effect of covariates is multiplicative, but on the odds of the survival function instead of the hazard function. Let $H(t) = -\operatorname{logit}[S_0(t)]$. If $H(t)$ is strictly increasing, model (1.5) can be equivalently written as

$$H(T) = -\boldsymbol{Z}'\boldsymbol{\beta} + W\,, \tag{1.6}$$

where the random variable W follows a standard logistic distribution.

For the two-sample situation with $Z = 0$ or 1, the proportional odds model says that the odds of the survival between the two samples are proportional to each other. Let $S_0(t)$ denote the survival function for the subjects with $Z = 0$ and $\alpha = \exp(-\beta)$. Then under model (1.5), we have

$$\frac{\lambda(t; Z=1)}{\lambda(t; Z=0)} = \frac{1}{1 + (\alpha - 1)\, S_0(t)}\,,$$

which is a monotonic function of t and converges to 1 as $t \to \infty$. In other words, unlike with the PH model, the ratio of the hazards changes with time and are not proportional to each other under model (1.5). For applications of model (1.5) to the analysis of right-censored failure time data, see Chen (2001), Murphy et al. (1997), and Yang and Prentice (1999) among others.

1.4.5 The Additive Hazards Model

As with the PH model, the additive hazards model specifies the effect of covariates on the failure time through the hazard function. Specifically, it assumes that given $\boldsymbol{Z}$, the hazard function of T has the additive form given by

$$\lambda(t; \boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{Z}'\,\boldsymbol{\beta}\,, \tag{1.7}$$

where $\lambda_0(t)$ is an arbitrary unspecified baseline hazard function, and $\boldsymbol{\beta}$ represents covariate effect as before. That is, the effect of covariates is to additively increase or decrease the hazard function.

Although both the PH model and the additive hazards model focus on the hazard function, the defined covariate effects have different meanings. Under the PH model, the regression parameter $\boldsymbol{\beta}$ represents the logrithm of the risk ratio in terms of risk factors and failure rates, whereas under model (1.7), $\boldsymbol{\beta}$ denotes the risk difference. This easily can be seen in the two-sample situation where Z takes only value 0 or 1. In this case, we have

$$\lambda(t; Z=1) = \lambda(t; Z=0) + \beta\,.$$

One attractive feature of the additive hazards model is that it provides a simple structure for modeling failure time data when there exist latent variables or frailties. For the additive frailty model, the marginal model is still the additive hazards model and the regression parameter $\boldsymbol{\beta}$ has the same meaning in both the the additive frailty model and the marginal model (Lin, Oakes and Ying, 1998; Lin and Ying, 1997). The same is not true for the PH model.

The authors who discussed model (1.7) include Aalen (1980), Breslow and Day (1987), Kim and Lee (1998), Kulich and Lin (2000), and Lin and Ying (1994). Some generalizations of model (1.7) have been proposed to make it more flexible for the analysis of right-censored failure time data. For example, we could allow $\boldsymbol{Z}$ to be time-dependent, and in this case, inferences about $\boldsymbol{\beta}$ are similar to those for the situation where $\boldsymbol{Z}$ is time-independent. Lin and Ying (1995) gave an additive-multiplicative hazard model that combines the PH model and the additive hazard model together. Martinussen and Scheike (2002a) and Scheike and Zhang (2002) further generalized the model.

1.4.6 The Accelerated Failure Time Model

The accelerated failure time model specifies that

$$\log T \;=\; \boldsymbol{Z}'\boldsymbol{\beta} + W \,, \tag{1.8}$$

where $\boldsymbol{\beta}$ is defined as before, and W is an error variable with an unknown distribution function. It is interesting to note that the ways that covariates affect the failure time in the accelerated failure time model and the proportional odds model are similar, but the ways that they affect the survival function are quite different as seen below.

Define $W^* = \exp(W)$ and let $\lambda_w(t)$ denote the hazard function of W^*, which is independent of $\boldsymbol{\beta}$. Then $T = \exp(\boldsymbol{Z}'\boldsymbol{\beta})\,W^*$, and the hazard and survival functions of T given $\boldsymbol{Z}$ have the forms

$$\lambda(t;\boldsymbol{Z}) \;=\; \lambda_w(t\,e^{-\boldsymbol{Z}'\boldsymbol{\beta}})\,\exp(-\boldsymbol{Z}'\boldsymbol{\beta})$$

and

$$S(t;\boldsymbol{Z}) \;=\; \exp\left[-\Lambda_w(t\,e^{-\boldsymbol{Z}'\boldsymbol{\beta}})\right]\,,$$

respectively, where $\Lambda_w(t) = \int_0^t \lambda_w(s)ds$.

It is interesting to note that under model (1.8), the effect of covariates is also multiplicative as under the PH model, but on t instead of the hazard function. In other words, the effect is to change the timescale and therefore to accelerate or decelerate the time to failure. Although the PH model specifies that the effect of covariates on the hazard is multiplicative, it does not give a direct relationship between $\boldsymbol{Z}$ and T because $\lambda_0(t)$ is arbitrary. In contrast, the model (1.8) specifies a linear relationship between $\log T$ and $\boldsymbol{Z}$.

Consider the two-sample situation where $Z = 0$ or 1 and let $S_1(t)$ and $\lambda_1(t)$ denote the survival and hazard functions of the subjects with $Z = 1$, respectively. Then we have

$$S_1(t) \;=\; S_0(\gamma\,t) \;\;, \lambda_1(t) \;=\; \gamma\,\lambda_0(\gamma\,t)\,,$$

where $\gamma = \exp(-Z'\beta)$ and $S_0(t)$ and $\lambda_0(t)$ are the survival and hazard functions of the subjects with $Z = 0$, respectively.

As for the PH model, one could also consider model (1.8) with time-dependent covariates. References that discuss the application of model (1.8) for the analysis of right-censored failure time data include Bagdonavicius and Nikulin (2001), Jin et al. (2003), Kalbfleisch and Prentice (2002), Park and Wei (2003), Tsiatis (1990), Wei et al. (1990), and Ying (1990).

1.4.7 The Linear Transformation Model

All regression models described above are specific models in the sense that they are single models and the underlying probability distribution is completely specified given unknown baseline functions. This subsection introduces a class of regression models, commonly called the linear transform model. Let T and $\boldsymbol{Z}$ be defined as before and $H(t)$ an unknown strictly increasing function. A linear transform model specifies that

$$H(T) \;=\; \boldsymbol{Z}'\boldsymbol{\beta} + W \,, \tag{1.9}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters as before, and the random variable W has a completely known distribution function F.

From (1.6), (1.8), and (1.9), it is seen that the linear transformation model clearly has close relationships with the accelerated failure time model and the proportional odds model. Under them, the ways by which explanatory variables affect the failure time are similar although details may be different. Furthermore, model (1.9) actually includes the PH model and the proportional odds model as special cases. To obtain the PH model, one can take $F(t) = 1 - \exp[-\exp(t)]$, the extreme value distribution, in model (1.9). The linear transformation model gives the proportional odds model if we let F be the standard logistic distribution.

Equivalently, the linear transformation model can be defined by

$$g[S(t;\boldsymbol{Z})] \;=\; H(t) \;-\; \boldsymbol{Z}'\boldsymbol{\beta} \,,$$

where $g^{-1}(s) \;=\; 1 \;-\; F(s)$. This equation shows that under model (1.9), the effect of covariates is to shift the location of the survival function in the scale of g. A major advantage of the linear transformation model is its generality as F could be any distribution function. References that discuss the application of model (1.9) to regression analysis of right-censored failure time data include Chen et al. (2002), Cheng et al. (1995, 1997), Fine et al. (1998), Kong et al. (2004), and Lu and Ying (2004).

1.4.8 Discrete Regression Models

All regression models discussed so far are for continuous survival variables. This subsection presents two commonly used regression models for discrete survival variables. One is the grouped PH model (Pierce et al., 1979; Prentice and Gloeckler, 1978) and the other is the logistic model (Lawless, 2003). Using the notation defined in the previous subsections, the grouped PH model assumes that given $\boldsymbol{Z}$, the survival function at time s_j has the form

$$S(s_j;\boldsymbol{Z}) \,=\, [S_0(s_j)]^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}\,. \tag{1.10}$$

In this model, $S_0(s_j)$ denotes the value of the baseline survival function or survival function for the subjects with $\boldsymbol{Z} \,=\, 0$ at s_j and $\boldsymbol{\beta}$ regression parameters. This gives

$$q_j(\boldsymbol{Z}) \,=\, P(T > s_j | T \geq s_j, \boldsymbol{Z}) \,=\, q_j^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}$$

and

$$S(s_j;\boldsymbol{Z}) \,=\, \prod_{k=1}^{j} q_k^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}\,,$$

where $q_j \,=\, S_0(s_j)/S_0(s_{j-1})$.

In practice, it is common to reparameterize model (1.10) by taking $\alpha_j \,=\, \log(-\log q_j)$. This not only removes the rang restriction on parameters but also improves the convergence in the determination of parameter estimates. Using the new parameters, we have

$$S(s_j;\boldsymbol{Z}) \,=\, \prod_{k=1}^{j} e^{-\,\exp(\alpha_k \,+\, \boldsymbol{Z}'\boldsymbol{\beta})}\,. \tag{1.11}$$

Model (1.10) can be regarded as arising from the PH model (1.4) due to the grouping of continuous failure times. In this case, it is supposed that each subject can be possibly observed only at the s_j's and only values of the survival function at these time points are of interest or can be estimated. Among others, Lawless (2003) and Prentice and Gloeckler (1978) discussed use of the grouped PH model for regression analysis of right-censored grouped failure time data.

In contrast with model (1.10), the logistic model cannot be obtained by grouping from the PH model, and it specifies that

$$q_j(\boldsymbol{Z}) \,=\, P(T > s_j | T \geq s_j, \boldsymbol{Z}) \,=\, \frac{1}{1 \,+\, \gamma_j\, e^{\boldsymbol{Z}'\boldsymbol{\beta}}}\,,$$

where

$$\gamma_j \,=\, \frac{1 - q_j(\boldsymbol{0})}{q_j(\boldsymbol{0})} \,=\, \frac{1 - q_j}{q_j}\,.$$

Then the conditional survival function given $\boldsymbol{Z}$ has the form

$$S(s_j;\boldsymbol{Z}) \,=\, \prod_{k=1}^{j} (1 \,+\, \gamma_k\, e^{\boldsymbol{Z}'\boldsymbol{\beta}})^{-1}\,. \tag{1.12}$$

The logistic model can be equivalently defined by

$$\log\left[\frac{1-q_j(\boldsymbol{Z})}{q_j(\boldsymbol{Z})}\right] = \log\gamma_j + \boldsymbol{Z}'\boldsymbol{\beta}\,.$$

For model (1.11), one may want to use the reparameterization $\alpha_j = \log\gamma_j$, which serves the same purpose as the reparameterization discussed above for the grouped PH model. The model (1.11) was initially proposed by Cox (1972) and developed further by Thompson (1977) for regression analysis of right-censored failure time data. An attractive feature of this model is that for inference about regression parameter $\boldsymbol{\beta}$, a partial likelihood for $\boldsymbol{\beta}$ can be derived and used, which does not involve the baseline survival function or the γ_j's (Lawless, 2003). The same is not possible for the grouped PH model.

2

Inference for Parametric Models and Imputation Approaches

2.1 Introduction

Although the main focus of this book is nonparametric and semiparametric inference procedures, it is helpful to first consider inference methods for parametric models and some imputation approaches. A main advantage of parametric approaches is that their implementation is straightforward in principle and in fact standard maximum likelihood theory generally applies. Imputation approaches are used to reduce the problem of analyzing interval-censored failure time data to that of analyzing right-censored failure time data. Thus one can avoid dealing with interval censoring and use existing inference procedures and statistical software developed for right-censored data.

Section 2.2 describes several commonly used parametric models for failure time variables with or without the existence of covariates. In Section 2.3, inference for these models is discussed with the focus on the standard likelihood-based inference procedures that generally apply to most parametric models. The imputation approach for the analysis of interval-censored failure time data is the topic of Section 2.4. Section 2.5 provides bibliographic notes and general discussion about parametric and imputation approaches.

2.2 Parametric Failure Time Models

This section describes several commonly used parametric models for T, a nonnegative random variable representing the failure time of a subject. These include the exponential model, Weibull model, log-normal model, and log-logistic model. Some other parametric models can be found in Kalbfleisch and Prentice (2002) and Lawless (2003).

2.2.1 The Exponential Model

The one-parameter exponential model assumes that the hazard function of T is constant over the range of T. That is,

$$\lambda(t) = \lambda > 0 .$$

It is the simplest failure time model and supposes that the instantaneous failure rate is independent of time t. Under this model, the survival and density functions of T are, respectively,

$$S(t) = e^{-\lambda t} , \quad f(t) = \lambda e^{-\lambda t} .$$

Furthermore, it can be easily shown that the conditional probability of failure within a time interval of specified length is the same regardless of how long the subject has been on study. This property is usually referred to as the memoryless property of the exponential model.

Suppose that there exists a vector of covariates $\boldsymbol{Z}$ and one is interested in the effect of $\boldsymbol{Z}$ on T. One way to define an exponential regression model is to assume that the conditional hazard function of T given $\boldsymbol{Z}$ has the form

$$\lambda(t; \boldsymbol{Z}) = \lambda \exp(\boldsymbol{Z}' \boldsymbol{\beta})$$

that follows the PH model (1.4). Here $\boldsymbol{\beta}$ denotes the vector of regression parameters. The conditional density function of T then has the form

$$f(t; \boldsymbol{Z}) = \lambda \exp(\boldsymbol{Z}' \boldsymbol{\beta}) \exp[-\lambda t \exp(\boldsymbol{Z}' \boldsymbol{\beta})]$$

given $\boldsymbol{Z}$.

Define $Y = \log T$, the log survival time. Then the model above can be equivalently defined by

$$Y = \alpha - \boldsymbol{Z}' \boldsymbol{\beta} + W , \tag{2.1}$$

where $\alpha = -\log \lambda$ and W has the extreme value distribution with the density function given by

$$\exp(w - e^{w}) , \ -\infty < w < \infty .$$

With respect to T, model (2.1) is a log-linear model, and for Y, it is a linear model with the error variable W having the extreme value distribution.

2.2.2 The Weibull Model

The simple exponential model described above depends only on one parameter and can be too restrictive sometimes. An important generalization of it is the two-parameter Weibull model with the hazard function

$$\lambda(t) = \lambda \gamma (\lambda t)^{\gamma - 1}$$

for $\lambda , \gamma > 0$. It is easy to see that this hazard function is monotone decreasing for $\gamma < 1$, increasing for $\gamma > 1$, and reduces to the exponential hazard if $\gamma = 1$. Under the Weibull model, the survival and density functions of T have the forms

$$S(t) = \exp[-(\lambda t)^{\gamma}]$$

and

$$f(t) = \lambda \gamma (\lambda t)^{\gamma-1} \exp[-(\lambda t)^{\gamma}] ,$$

respectively.

For regression analysis, as in the case of exponential model, the hazard function can be generalized to

$$\lambda(t; \boldsymbol{Z}) = \lambda \gamma (\lambda t)^{\gamma-1} \exp(\boldsymbol{Z}' \boldsymbol{\beta}) .$$

The corresponding conditional density function of T given $\boldsymbol{Z}$ is then

$$f(t; \boldsymbol{Z}) = \lambda \gamma (\lambda t)^{\gamma-1} \exp(\boldsymbol{Z}' \boldsymbol{\beta}) \exp[-(\lambda t)^{\gamma} \exp(\boldsymbol{Z}' \boldsymbol{\beta})] .$$

As in the case of the exponential model, with $Y = \log T$, this regression model can be written as

$$Y = \alpha + \boldsymbol{Z}' \boldsymbol{\beta}^* + \sigma W , \tag{2.2}$$

where $\alpha = -\log \lambda$, $\sigma = \gamma^{-1}$, $\boldsymbol{\beta}^* = -\sigma \boldsymbol{\beta}$, and W follows the extreme value distribution.

It is interesting to note that as the PH model, both exponential and Weibull regression models specify that covariates have multiplicative effects on the hazard function. On the other hand, like the accelerated failure time model (1.8), both models are log-linear models and under them, covariates additively affect the log survival time Y. In fact, the Weibull model is the only family of models satisfying these conditions (Kalbfleisch and Prentice, 2002).

2.2.3 The Log-normal Model

The log-normal model assumes that the log survival time $Y = \log T$ has the form $Y = \alpha + \sigma W$ with W being a standard normal variable. The density function of T is then

$$f(t) = (2\pi)^{-1/2} \gamma t^{-1} \exp\left[\frac{-\gamma^2 (\log \lambda t)^2}{2} \right] ,$$

where $\lambda = \exp(-\alpha)$ and $\gamma = \sigma^{-1}$ as before. The survival and hazard functions of T involve the standard normal distribution function $\Phi(w)$ with

$$S(t) = 1 - \Phi(\gamma \log \lambda t)$$

and both have no closed form. The hazard function increases from zero at $t = 0$ to a maximum and then decreases to zero as t increases.

In the case when there exist covariates $\boldsymbol{Z}$, it is apparent that following models (2.1) and (2.2), one can define the log-normal regression model as

$$Y = \alpha + \boldsymbol{Z}' \boldsymbol{\beta} + \sigma W , \tag{2.3}$$

the usual linear regression model. This model is particularly easy to apply if there exists no censoring. But with censoring, the computation and inference become difficult.

2.2.4 The Log-logistic Model

The log-logistic model is defined in the same way as the log-normal model except that W has the logistic density

$$\frac{e^w}{(1 + e^w)^2} \, .$$

This density function is symmetric with mean 0 and variance $\pi^2/3$, having slightly heavier tails than the normal density function. Under the log-logistic model, T has the density function

$$f(t) \, = \, \lambda \, \gamma \, (\lambda \, t)^{\gamma - 1} \, [1 \, + \, (\lambda \, t)^{\gamma}]^{-2} \, ,$$

where again $\lambda \, = \, \exp(-\alpha)$ and $\gamma \, = \, \sigma^{-1}$.

In comparison with the log-normal model, the log-logistic model, although it is used less frequently in failure time analysis, has the advantage that both survival and hazard functions have closed forms. Thus it is more convenient than the log-normal model in handling censoring. The survival and hazard functions are, respectively,

$$S(t) \, = \, \frac{1}{1 \, + \, (\lambda \, t)^{\gamma}}$$

and

$$\lambda(t) \, = \, \frac{\lambda \, \gamma \, (\lambda \, t)^{\gamma - 1}}{1 \, + \, (\lambda \, t)^{\gamma}} \, .$$

If $\gamma \, < \, 1$, $\lambda(t)$ is monotone decreasing from ∞ and if $\gamma \, = \, 1$, it is monotone decreasing from λ. For $\gamma \, > \, 1$, as the log-normal hazard function, $\lambda(t)$ increases from zero to a maximum and then decreases to zero.

2.3 Likelihood-based Inference for Parametric Models

Consider a survival study that consists of n independent subjects. Let T_i denote the survival time of interest for subject i, $i \, = \, 1, ..., n$, and suppose that the T_i's follow a parametric model with survival function $S(t, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \, = \, (\theta_1, ..., \theta_p)'$ denotes unknown parameters. Also suppose that only interval-censored data are available and they have the form

$$\{ \, (L_i, R_i], \boldsymbol{Z}_i \, ; \, i \, = \, 1, ..., n \, \} \, ,$$

where $(L_i, R_i]$ denotes the interval to which T_i is observed to belong and $\boldsymbol{Z}_i$ is the covariate vector associated with subject i, $i \, = \, 1, ..., n$. Then the likelihood function is proportional to

$$L(\boldsymbol{\theta}) \, = \, \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) \, = \, \prod_{i=1}^{n} [\, S(L_i, \boldsymbol{\theta}) \, - \, S(R_i, \boldsymbol{\theta}) \,]$$

assuming that $L_i < R_i$ for all $i = 1, ..., n$.

In the following, general likelihood-based inference procedures are discussed first and followed by inference procedures about exponential and general log-linear regression models. Two examples are then provided.

2.3.1 Inference with General Parametric Models

A standard approach to inferences about $\boldsymbol{\theta}$ is to estimate it by the maximum likelihood estimator, defined as the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, and to use the score and likelihood ratio statistics derived from $L(\boldsymbol{\theta})$. Given the independent censoring mechanism assumed here, standard large-sample likelihood theory generally applies. In particular, asymptotic approximations to the distributions of the maximum likelihood estimator, score statistic, and likelihood ratio statistic are available and provide straightforward inference approaches.

Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$. Define

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} L_i(\boldsymbol{\theta})$$

and

$$I(\boldsymbol{\theta}) = \sum_{i=1}^{n} I_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} L_i(\boldsymbol{\theta}) \, .$$

Then under certain regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent. Furthermore, when n is large, it is the unique solution to $U(\boldsymbol{\theta}) = 0$, and its distribution can be approximated by the multivariate normal distribution with mean θ and the covariance matrix $I^{-1}(\boldsymbol{\theta})$. In other words, one has

$$\hat{\boldsymbol{\theta}} \sim N(\, \boldsymbol{\theta} \, , \, I^{-1}(\boldsymbol{\theta}) \,) \, ,$$

which can be used for testing hypotheses and deriving interval estimates for $\boldsymbol{\theta}$. To determine $\hat{\boldsymbol{\theta}}$, one can use any root-finding procedure or the Newton-Raphson algorithm. Suppose that $I(\boldsymbol{\theta}_0)$ is nonsingular. In the Newton-Raphson algorithm, an initial value, say $\boldsymbol{\theta}_0$, of $\boldsymbol{\theta}$ is updated by

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - I^{-1}(\boldsymbol{\theta}_0) \, U(\boldsymbol{\theta}_0)$$

iteratively until convergence is achieved.

To test the hypothesis $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is known, it is convenient to use the score test statistic

$$U'(\boldsymbol{\theta}_0) \, I^{-1}(\boldsymbol{\theta}_0) \, U(\boldsymbol{\theta}_0) \, ,$$

which has an asymptotic χ^2 distribution with degrees of freedom p. Assume $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are components of $\boldsymbol{\theta}$ with dimensions k and $p-k$, respectively. Then in practice, a more common hypothesis is $H_2 : \boldsymbol{\theta}_1 =$

$\boldsymbol{\theta}_{10}$, where $\boldsymbol{\theta}_{10}$ is known. In this situation, partition $U(\boldsymbol{\theta})$ in the same way as $\boldsymbol{\theta}$, i.e.,

$$U'(\boldsymbol{\theta}) = [U_1'(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), U_2'(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] ,$$

where U_1 and U_2 are of dimensions k and $p-k$ corresponding with $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Let $\hat{\boldsymbol{\theta}}_{20}$ denote the estimate of $\boldsymbol{\theta}_2$ given by the solution to $U_2(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2) = 0$ and $I^{11}(\boldsymbol{\theta})$ the submatrix of $I^{-1}(\boldsymbol{\theta})$ corresponding with $\boldsymbol{\theta}_1$. Then the test of H_2 can be based on the statistic

$$(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})' \, [I^{11}(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20})]^{-1} \, (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \tag{2.4}$$

or

$$U_1'(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20}) \, I^{11}(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20}) \, U_1(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20}) , \tag{2.5}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2')'$. Both statistics have an asymptotic χ^2 distribution with the degrees of freedom k.

In the methods described above, one needs to determine the observed Fisher information matrix $I(\boldsymbol{\theta})$, which could be difficult sometimes. An approach without this drawback is based on the likelihood ratio statistic

$$LR(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} .$$

One can use the statistic $-2\,\log\,[LR(\boldsymbol{\theta}_0)]$ or $-2\,\log\,[LR(\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}_{20})]$ to test the hypothesis H_1 or H_2, respectively. They have an asymptotic χ^2 distribution with the degrees of freedom p and k, respectively, under the corresponding hypothesis.

2.3.2 Inference with the Exponential Regression Model

This subsection discusses a special situation where T_i follows the exponential regression model given in Section 2.2.1. For ease of notation, suppose that the hazard function has the form

$$\lambda(t\,;\,\boldsymbol{Z}_i) = \exp(\boldsymbol{Z}_i'\,\boldsymbol{\beta}) ,$$

where $Z_{i1} = 1$. With this notation, one has that $\boldsymbol{\theta} = \boldsymbol{\beta} = (\beta_1, ..., \beta_{p+1})'$ and $\lambda = \exp(\beta_1)$, the hazard rate when all real covariates have value 0.

For subject i, define $\delta_i = 1$ if $L_i = R_i$ and $\delta_i = 0$, otherwise $i = 1, ..., n$. That is, δ_i indicates if an exact observation is observed for subject i. Then the likelihood function $L(\boldsymbol{\theta})$ is

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \exp(-e^{\boldsymbol{Z}_i'\boldsymbol{\beta}}\,L_i) \prod_{i:\delta_i=0} \left[\exp(-e^{\boldsymbol{Z}_i'\boldsymbol{\beta}}\,L_i) - \exp(-e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} R_i)\right]$$

conditional on the $\boldsymbol{Z}_i$'s. The score vector and the observed Fisher information matrix are

$$U(\boldsymbol{\beta}) = \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n} \boldsymbol{Z}_i \left[\delta_i (1 - e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} L_i) + (1 - \delta_i) \exp(\boldsymbol{Z}_i' \boldsymbol{\beta}) \, d_i(\boldsymbol{\beta}, \boldsymbol{Z}_i) \right]$$

and

$$\hat{I}(\boldsymbol{\beta}) = - \frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$$

$$= \sum_{i=1}^{n} \boldsymbol{Z}_i \, \boldsymbol{Z}_i' \, e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} \left\{ \delta_i \, L_i - (1 - \delta_i) \left[d_i(\boldsymbol{\beta}, \boldsymbol{Z}_i) + e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} \, d_i^*(\boldsymbol{\beta}, \boldsymbol{Z}_i) \right] \right\} ,$$

respectively, where

$$d_i(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \frac{R_i \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} R_i) - L_i \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} L_i)}{\exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} L_i) - \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} R_i)}$$

and

$$d_i^*(\boldsymbol{\beta}, \boldsymbol{Z}_i) = \frac{L_i^2 \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} L_i) - R_i^2 \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} R_i)}{\exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} L_i) - \exp(-e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} R_i)} - d_i^2(\boldsymbol{\beta}, \boldsymbol{Z}_i) \, .$$

For the two sample comparison problem, one has that $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})'$, where $Z_{i1} = 1$ as before and $Z_{i2} = 0$ or 1. The comparison is equivalent to testing $\beta_2 = 0$ and can be performed by using statistic (2.4) or (2.5) with $\theta_1 = \beta_2$ and $\theta_2 = \beta_1$.

2.3.3 Inference with Log-linear Regression Models

This subsection considers the more general situation where instead of the exponential model, T_i follows the general log-linear model including those defined in (2.1) to (2.3). As before, define $Y_i = \log(T_i)$ and suppose that the density function of the Y_i's is given by

$$\sigma^{-1} f(w) \, ,$$

where $w = (y - \boldsymbol{Z}' \boldsymbol{\beta})/\sigma$. Here again we assume that $Z_{i1} = 1$ for all i. Then the likelihood function of $\boldsymbol{\beta}$ and σ can be written as

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \left[\sigma^{-1} f(w_{L_i}) \right]^{\delta_i} \left[S(w_{L_i}) - S(w_{R_i}) \right]^{1-\delta_i} ,$$

where $w_{L_i} = (\log L_i - \boldsymbol{Z}_i' \boldsymbol{\beta})/\sigma$, $w_{R_i} = (\log R_i - \boldsymbol{Z}_i' \boldsymbol{\beta})/\sigma$, $S(w) = \int_w^\infty f(s) \, ds$, and the δ_i's are defined as before.

The score vector has the form

$$\begin{aligned} U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\sigma) &= \frac{\partial \log\, L(\boldsymbol{\beta},\sigma)}{\partial \boldsymbol{\beta}} \\ &= -\,\sigma^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i \left[\delta_i\,\frac{f'(w_{L_i})}{f(w_{L_i})} \,-\, (1-\delta_i)\,\frac{f(w_{L_i})-f(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right] \end{aligned}$$

and

$$\begin{aligned} U_{\sigma}(\boldsymbol{\beta},\sigma) &= \frac{\partial \log\, L(\boldsymbol{\beta},\sigma)}{\partial \sigma} \\ = \,-\sigma^{-1}\sum_{i=1}^{n} &\left[\delta_i \,+\, \delta_i\,\frac{w_{L_i}\,f'(w_{L_i})}{f(w_{L_i})} \,-\, (1-\delta_i)\,\frac{w_{L_i}\,f(w_{L_i})-w_{R_i}\,f(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right], \end{aligned}$$

where $f'(w) \,=\, df(w)/dw$. The components of the observed Fisher information matrix are

$$\begin{aligned} -\,\frac{\partial^2 \log\, L(\boldsymbol{\beta},\sigma)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} &= \sigma^{-2}\sum_{i=1}^{n} \boldsymbol{Z}_i\,\boldsymbol{Z}_i'\,\left\{\delta_i\,\left[\left(\frac{f'(w_{L_i})}{f(w_{L_i})}\right)^2 \,-\, \frac{f''(w_{L_i})}{f(w_{L_i})}\right]\right. \\ &\left.+\,(1-\delta_i)\,\left[\left(\frac{f(w_{L_i})-f(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right)^2 \,+\, \frac{f'(w_{L_i})-f'(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right]\right\}, \end{aligned}$$

$$\begin{aligned} -\,\frac{\partial^2 \log\, L(\boldsymbol{\beta},\sigma)}{\partial\boldsymbol{\beta}\partial\sigma} &= \sigma^{-2}\sum_{i=1}^{n} \boldsymbol{Z}_i\,\left\{\delta_i\,w_{L_i}\,\left[\left(\frac{f'(w_{L_i})}{f(w_{L_i})}\right)^2 \,-\, \frac{f''(w_{L_i})}{f(w_{L_i})}\right]\right. \\ &+\,(1-\delta_i)\,\left[\frac{(f(w_{L_i})-f(w_{R_i}))\,(w_{L_i}f(w_{L_i})-w_{R_i}f(w_{R_i}))}{(S(w_{L_i})-S(w_{R_i}))^2}\right. \\ &\left.\left.+\,\frac{w_{L_i}f'(w_{L_i})-w_{R_i}f'(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right]\right\} \,+\, \sigma^{-1}\,U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\sigma)\,, \end{aligned}$$

and

$$\begin{aligned} -\,\frac{\partial^2 \log\, L(\boldsymbol{\beta},\sigma)}{\partial\sigma\partial\sigma} &= \sigma^{-2}\sum_{i=1}^{n}\left\{\delta_i\,\left[\left(\frac{w_{L_i}f'(w_{L_i})}{f(w_{L_i})}\right)^2\right.\right. \\ -\,\frac{w_{L_i}f'(w_{L_i})+w_{L_i}^2f''(w_{L_i})}{f(w_{L_i})}&\left.\right] \,+\, (1-\delta_i)\,\left[\left(\frac{w_{L_i}f(w_{L_i})-w_{R_i}f(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}\right)^2\right. \\ +\,\frac{w_{L_i}f(w_{L_i})+w_{L_i}^2f'(w_{L_i})-w_{R_i}f(w_{R_i})-w_{R_i}^2f'(w_{R_i})}{S(w_{L_i})-S(w_{R_i})}&\left.\left.\right]\right\} \,+\, \sigma^{-1}\,U_{\sigma}(\boldsymbol{\beta},\sigma)\,, \end{aligned}$$

where $f''(w) \,=\, d^2f(w)/dw^2$.

Several authors have investigated parametric inference procedures for the analysis of interval-censored data. For example, two recent references are Lindsey (1998) and Lindsey and Ryan (1998). The former considered a number of commonly used parametric models and suggested that for many situations, one can use the middle point imputation approach, which is described in the next section. The latter gave a useful tutorial about parametric approaches as well as nonparametric approaches and in particular, discussed the log-linear regression models considered above. More inferences are given in Section 2.5.

Table 2.1. Observed intervals in months for HIV infection times of 297 Danish homosexuals with 0 denoting December 1979 and n_i being multiplicities

L_i	R_i	n_i	L_i	R_i	n_i	L_i	R_i	n_i
0	24	24	0	39	2	24	28	4
24	39	1	24	57	10	24	88	3
24	113	4	24	∞	61	28	39	4
28	88	1	28	∞	8	39	57	3
39	113	2	39	∞	15	57	88	5
57	113	1	57	∞	22	88	113	1
88	∞	34	113	∞	92			

2.3.4 Two Examples

To illustrate the inference approaches discussed in the previous subsections, we discuss two examples. First, consider the set of interval-censored data presented in Table 2.1, reproduced from Carstensen (1996). The data concern HIV infection times of 297 Danish homosexuals, who were supposed to be examined for their HIV status at six time points between December 1981 and May 1989. As expected, many patients did not make all six visits, and as seen from the table, the data are highly interval-censored. In the table, zero means December 1979, and the time origin is assumed to be the same for all patients.

For the analysis, we first fit the data to the exponential model described in Section 2.2.1, and the likelihood-based approach gives $\hat{\lambda} = 0.0034$ with an estimated standard error of 0.0004. If we use the Weibull model given in Section 2.2.2, the approach gives $\hat{\lambda} = 0.0025$ and $\hat{\gamma} = 0.8119$ with their estimated standard errors equal to 0.0002 and 0.2523, respectively. In practice, one use of the Weibull model is to test if the exponential model is appropriate for the data. This can be carried out by testing $\gamma = 1$, for which one has the Wald statistic equal to $(0.8119 - 1)/0.2523 = -0.7455$. Based on the standard normal distribution, this gives a p-value of 0.4560 and suggests that the exponential model seems to provide a reasonable fit to the data. For the data set, one question of interest is to estimate the proportion of the patients who were HIV-positive by 1990, which corresponds with the probability $P_{120} = P(\,T \leq 120\,) = 1 - S(120)$. For this, Figure 2.1 displays the estimated proportion curves, or the cumulative probabilities of a patient being HIV-positive by time t under exponential and Weibull models, respectively. It can be seen that using the exponential model, one gets $\hat{P}_{120} = 0.3350$, suggesting that about 34% of the patients would be HIV-positive by 1990. Under the Weibull model, one obtains $\hat{P}_{120} = 0.3136$, that is, the proportion would be about 31%.

For the second example, we discuss the analysis of the interval-censored failure time data given in Table 1.4 from an early breast cancer study. The study consists of two treatments, radiation therapy alone and radiation therapy plus adjuvant chemotherapy, and its main goal is to compare the two treat-

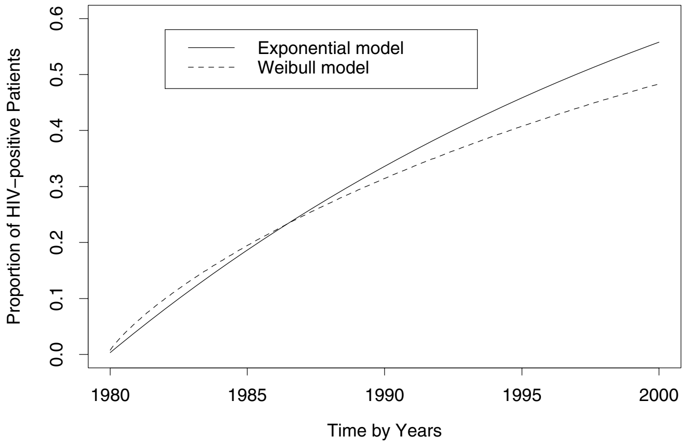

Fig. 2.1. Estimated proportion curves of the patients being HIV-positive by certain time.

ments in terms of time to breast retraction. However, only interval-censored observations on the time to breast retraction are available.

For the comparison, we define $Z_i = 0$ for the patients given radiation therapy alone and $Z_i = 1$ otherwise. The fit of the exponential regression model then gives $\hat{\lambda} = 0.0163$ and $\hat{\beta} = 0.7416$ with their estimated standard errors being 0.0036 and 0.2769, respectively, using the notation in Section 2.2.1. Note that the comparison of the two treatments is equivalent to testing $\beta = 0$, for which the Wald test yields $\hat{\beta}/0.2769 = 2.6782$ and gives a p-value of 0.0074 based on the standard normal distribution. This suggests that the patients given radiation therapy plus adjuvant chemotherapy have significantly higher risk to develop breast retraction. In other words, the adjuvant chemotherapy significantly increases the risk of breast retraction. Using the Weibull regression model in Section 2.2.2, we get $\hat{\lambda} = 0.0203$, $\hat{\gamma} = 1.6149$, and $\hat{\beta} = 0.5677$ and the estimated standard errors are 0.0028, 0.1936, and 0.1757, respectively. In this case, we obtain a p-value of 0.0012 for testing $\beta = 0$.

2.4 Imputation-based Inference

Imputation or multiple imputation is a general approach for handling missing data problems (Rubin, 1987) and is commonly used in, for example, sample surveys. Missing data usually refer to observed data in which for some subjects, no information is observed for response variables of interest. Censored or interval-censored failure time data differ from missing data in nature because the former provides some incomplete information about failure variables of interest. In other words, interval-censored data are really incomplete data, not exactly missing data Nevertheless, one can still treat the underlying, unobserved true interval-censored failure times as missing and replace them by using some imputed times conditional on the observed information.

As in the previous section, let the T_i's represent the survival times of interest from n independent individuals, and suppose that only interval-censored data

$$\{ (L_i, R_i], \boldsymbol{Z}_i \,;\, i \,=\, 1, ..., n \}$$

are observed. For one sample problem, it is assumed that $\boldsymbol{Z}_i \,=\, 0$. Let $S(t\,;\,\boldsymbol{\theta})$ denote the survival function of the T_i's that is known up to unknown parameters $\boldsymbol{\theta}$. In the following, the discussion mainly focuses on the situation where the dimension of $\boldsymbol{\theta}$ is infinite. These include situations such that $\boldsymbol{\theta}$ represents the whole survival function, or consists of a finite-dimensional vector of parameters of interest plus a nuisance function. One sample nonparametric problem corresponds with the former situation, and an example of the latter situation is given by $\boldsymbol{\theta}$ being $\boldsymbol{\beta}$ and $\lambda_0(t)$ for regression analysis under model (1.4). However, as those discussed in the previous two sections, the methods described below equally apply to the case of finite-dimensional $\boldsymbol{\theta}$.

Imputation, in these cases, means to generate one or multiple sets of right-censored failure time data for the T_i's using the observed data. One then uses these new data to make inference about $\boldsymbol{\theta}$. It is apparent that instead of right-censored data, one could generate exact failure time data for the T_i's. However, this is usually not necessary and also not preferred. The main reason is that there exist many established methods for right-censored data for various inference problems and there are possible shortcomings of imputation approaches. In the following, two general imputation approaches are discussed. One is a single point imputation approach, which is commonly used in practice for its simplicity. The other is a multiple imputation approach (Wei and Tanner, 1991), the application of the data augmentation technique discussed in Tanner and Wong (1987) and Tanner (1991).

2.4.1 A Single Point Imputation Approach

For interval-censored failure time data, the simplest imputation approach is perhaps to assume that for subject i, the underlying true failure time T_i is equal to a value within the observed interval $(L_i, R_i]$, $i \,=\, 1, ..., n$. One common choice is to let T_i be the middle point of the interval for a finite interval or truly interval-censored observation. For intervals with $R_i \,=\, \infty$ or right-censored observations, the original observations are kept. Then we have a set of right-censored failure time data. An alternative to the mid-point imputation is to take T_i to be L_i, the left end point imputation, or R_i, the right end point imputation. It is apparent that these three methods would not give much different results if all finite intervals are narrow. In general, the selection can be made depending on if the true survival event under study is more likely to occur close to the middle, left end, or right end point of an observed interval. Of course, it is often the case that there does not exist any prior information about the part of an observed interval where the survival event is more likely to occur. In this case, rather than using the methods above, one can randomly select a value based on, for instance, the uniform distribution over the observed interval.

In the methods above, we only impute the T_i that are purely interval-censored. As mentioned before, this is partly because there exist many established inference approaches for right-censored data that can be applied to the imputed right-censored data. Suppose, for example, that one is interested in nonparametrically estimating a survival function. In this case, one can simply use the Kaplan-Meier estimator given by

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.6}$$

(Kalbfleisch and Prentice, 2002). Here the t_j's denote the distinct imputed exact failure times, and the d_j's and n_j's are the failure and risk numbers at each of the t_j's, respectively, based on the imputed right-censored data. It is apparent that $\hat{S}(t)$ is a step function with jumps and is discontinuous at the t_j's. The asymptotic variance of $\hat{S}(t)$ at time t can be estimated by

$$\hat{V}_S(t) = \hat{S}^2(t) \sum_{j:t_j \leq t} \frac{d_j}{n_j\,(n_j - d_j)} \tag{2.7}$$

(Kalbfleisch and Prentice, 2002), which is commonly referred to as Greenwood's formula (Greenwood, 1926).

Instead of the one sample problem, suppose that one is interested in regression analysis under the PH model (1.4). Let the t_j's be defined as above and $\boldsymbol{Z}_{(j)}$ denote the covariate vector of the subject whose imputed exact failure time is equal to t_j assuming that there are no tied imputed exact failure times. Also let $R(t_j)$ denote the risk set of individuals at time t_j based on the imputed right-censored data. Then the regression parameter $\boldsymbol{\beta}$ in model (1.4) can be estimated by the partial likelihood estimator defined as the value of $\boldsymbol{\beta}$ that maximizes the partial likelihood

$$L_p(\boldsymbol{\beta}) = \prod_j \frac{\exp(\boldsymbol{Z}'_{(j)}\,\boldsymbol{\beta})}{\sum_{l \in R(t_j)} \exp(\boldsymbol{Z}'_{(l)}\,\boldsymbol{\beta})} \tag{2.8}$$

(Cox, 1972). For right-censored failure time data, the partial likelihood estimator has been extensively studied and shown to be consistent and have an asymptotic multivariate normal distribution (Andersen, et al., 1982). If there exist tied imputed exact failure times, the partial likelihood $L_p(\boldsymbol{\beta})$ needs to be adjusted. For this, there exist several ways and one is to use the approximation

$$\prod_j \frac{\exp(\boldsymbol{Z}'_{(j)}\,\boldsymbol{\beta})}{[\,\sum_{l \in R(t_j)} \exp(\boldsymbol{Z}'_{(l)}\,\boldsymbol{\beta})\,]^{d_j}}$$

(Breslow, 1974), where d_j is the number of individuals whose imputed failure times are equal to t_j as above.

The biggest advantage of the single point imputation approach is its simplicity, and in this case, inference can be performed without any difficulty using existing software. If all intervals are relatively narrow or the overlapping among the intervals is slight, the approach can provide a reasonable and simple approximation to the inference based on observed data. In general, it may not be reliable, and clearly one natural extension is, instead of using it only once, to repeatedly carry out the approach as discussed below.

2.4.2 A Multiple Imputation Approach

This subsection discusses the application of the data augmentation algorithm given in Tanner and Wong (1987) and Tanner (1991) for the analysis of interval-censored data. The algorithm is originally designed to calculate the posterior distribution of the parameters of interest and iterates between two steps: imputation and posterior steps. In the former step, it first generates the parameter from the current estimate of its posterior distribution. Then M sets of unobserved (or complete) data are generated from the conditional distribution given the observed (or incomplete) data and the generated parameters. Here M is a prespecified integer. In the latter step, the posterior distribution is first obtained given the observed data and each set of unobserved data. The updated posterior distribution of the parameters is then given by the mixture of the M posterior distributions.

For the situations considered here, the interest is to make inference about $\boldsymbol{\theta}$ rather than a posterior distribution. Thus instead of imputation and posterior steps, one can use the following imputation and estimation steps. Specifically, let M be defined as above. By following the data augmentation algorithm, one can estimate $\boldsymbol{\theta}$ as follows.

Step 0. Give an initial value $\hat{\boldsymbol{\theta}}^{(0)}$ and set $\hat{S}^{(0)}(t) = S(t\,;\hat{\boldsymbol{\theta}}^{(0)})$.

Step 1. At the lth iteration, for each k and i, if $R_i = \infty$, i.e., a right-censored observation is observed for T_i, define $T_i^{(k,l)} = L_i$ and $\delta_i^{(k,l)} = 0$, $k = 1, ..., M$, $i = 1, ..., n$. Otherwise, define $T_i^{(k,l)}$ to be a random number generated from $\hat{S}^{(l-1)}$ conditional on $T_i^{(k,l)} \in (L_i, R_i]$ and $\delta_i^{(k,l)} = 1$. This gives M sets of right-censored data

$$\{\,T_i^{(k,l)}, \delta_i^{(k,l)}, \boldsymbol{Z}_i\,;\, i = 1, ..., n\,\}\,, \tag{2.9}$$

$k = 1, ..., M$. Note that here the right-censoring indicators are always same.

Step 2. For each of the M sets of right-censored data generated in step 1, obtain an estimate $\hat{\boldsymbol{\theta}}^{(k,l)}$. Suppose that we can also obtain an estimate of the covariance matrix of the estimate of a finite-dimensional component of $\boldsymbol{\theta}$ (parameters of interest) or the pointwise variance of $\hat{\boldsymbol{\theta}}^{(k,l)}$ for the one sample nonparametric problem. For both cases, let $\hat{\Sigma}^{(k,l)}$ denote the estimate for simplicity.

Step 3. Determine the updated estimator of $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}^{(l)} = \frac{1}{M} \sum_{k=1}^{M} \hat{\boldsymbol{\theta}}^{(k,l)} .$$

The corresponding covariance matrix or variance function can be estimated by

$$\hat{\Sigma}^{(l)} = \frac{1}{M} \sum_{k=1}^{M} \hat{\Sigma}^{(k,l)} + \left(1 + \frac{1}{M}\right) \frac{\sum_{k=1}^{M} [\hat{\boldsymbol{\theta}}_t^{(k,l)} - \hat{\boldsymbol{\theta}}_t^{(l)}]\,[\hat{\boldsymbol{\theta}}_t^{(k,l)} - \hat{\boldsymbol{\theta}}_t^{(l)}]'}{M - 1} . \tag{2.10}$$

In (2.10), $\hat{\boldsymbol{\theta}}_t^{(k,l)}$ denotes the component of $\boldsymbol{\theta}$ that is of interest or the value of the function represented by $\boldsymbol{\theta}$ at time t for the one sample nonparametric problem.

Step 4. Repeat steps 1 to 3 until the desired convergence occurs.

In step 3, the variance estimate consists of two terms, representing the within-imputation (the first term) and between-imputation (the second term) estimation. The term $(1 + M^{-1})$ instead of one is used to take account of a finite number of imputations. For step 0, one simple way to obtain an initial value is to apply the single point imputation approach to the observed data and use the resulting estimate. For convergence of the algorithm, a natural approach is to put a criterion on the parameters of interest and to stop the iteration when the consecutive estimates of the parameters are close enough.

To illustrate the estimation algorithm given above, consider the one sample nonparametric problem with the goal of estimating a survival function. In this case, $\boldsymbol{\theta}$ represents a survival function and $\boldsymbol{Z}_i = 0$. To determine the initial estimate in step 0, one can use the Kaplan-Meier estimator (2.6) given by, for example, the mid-point imputation approach. In step 2, $\hat{\boldsymbol{\theta}}^{(k,l)}$ is again given by the estimator (2.6) based on the imputed right-censored data and $\hat{\Sigma}^{(k,l)}$ can be taken to be the variance estimate (2.7). The estimate (2.10) then becomes

$$\frac{1}{M} \sum_{k=1}^{M} \hat{V}_S^{(k,l)}(t) + \left(1 + \frac{1}{M}\right) \frac{\sum_{k=1}^{M} [\hat{S}^{(k,l)}(t) - \hat{S}^{(l)}(t)]^2}{M - 1} .$$

In the above, $\hat{S}^{(k,l)}(t)$ and $\hat{V}_S^{(k,l)}(t)$ denote the estimates given by (2.6) and (2.7), respectively, based on the imputed right-censored data (2.9) and

$$\hat{S}^{(l)}(t) = \frac{1}{M} \sum_{k=1}^{M} \hat{S}^{(k,l)}(t) .$$

As another example, consider the regression analysis problem under model (1.4). In this case, $\boldsymbol{\theta}$ consists of two parts: $\boldsymbol{\beta}$ and $S_0(t)$, the baseline survival function, and suppose that one is interested only in the regression parameter $\boldsymbol{\beta}$. For the initial estimate, one can take $\hat{\boldsymbol{\beta}}^{(0)}$ to be the partial likelihood estimator given by (2.8) based on the imputed right-censored data given by the mid point imputation approach. For $S_0(t)$, several estimators can be used, and one common choice is to take

$$\hat{S}_0^{(0)}(t) = \exp\left[- \sum_{j:t_j \leq t} \frac{d_j}{\sum_{l \in R(t_j)} \exp(\boldsymbol{Z}_l' \hat{\boldsymbol{\beta}}^{(0)})} \right]$$

based on the same imputed data (Kalbfleisch and Prentice, 2002). In this estimate, t_j, d_j, and $R(t_j)$ are defined as in (2.6) to (2.8). The same estimators can be used again in step 2, but based on data (2.9). For estimation of the covariance matrix of $\hat{\boldsymbol{\beta}}^{(k,l)}$, the observed Fisher information matrix from (2.8) can be used. For step 3, (2.10) has the form

$$\hat{\Sigma}^{(l)} = \frac{1}{M} \sum_{k=1}^{M} \hat{\Sigma}^{(k,l)} + \left(1 + \frac{1}{M}\right) \frac{\sum_{k=1}^{M} [\hat{\boldsymbol{\beta}}^{(k,l)} - \hat{\boldsymbol{\beta}}^{(l)}]\, [\hat{\boldsymbol{\beta}}^{(k,l)} - \hat{\boldsymbol{\beta}}^{(l)}]'}{M - 1} .$$

For both the one sample nonparametric and regression analysis problems discussed above, in the iterations, one only needs to consider estimates of survival or baseline survival function that put probability mass at distinct values of left and right end points of observed intervals. In other words, one can focus on discrete survival functions that jump at these distinct time points. The reason for this will be explained in Sections 3.2 and 3.3.

Several authors have considered the multiple imputation procedure described above or similar methods for special situations. For example, Bebchuk and Betensky (2000) discussed estimation of a hazard function based on interval-censored data. Pan (2000a) and Satten et al. (1998) studied the regression problem under the PH model for interval-censored data.

2.4.3 Two Examples

In this subsection, again we discuss the analysis of the breast cancer data considered in Section 2.3.4, but using the imputation approach. First, we consider estimation of the survival functions corresponding with the two treatments, radiation therapy alone (RT) and radiation therapy plus adjuvant chemotherapy (RCT), separately based on observed data within each treatment group. Figure 2.2 presents estimates of the two survival functions given by the left end point, mid-point, and right end point imputation approaches, respectively. The estimates given by the multiple imputation approach with $M = 30$, 50, and 100, respectively, are displayed in Figure 2.3. Both figures indicate that the patients in the RT treatment seem to have lower risk to develop breast

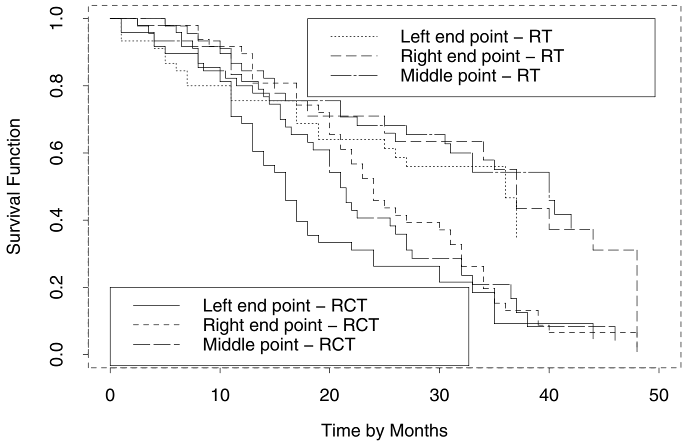

Fig. 2.2. Estimated survival functions using single imputation approaches.

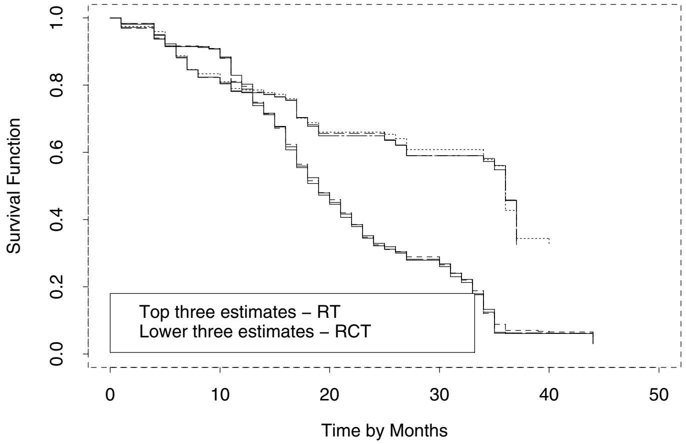

Fig. 2.3. Estimated survival functions using multiple imputation approaches.

retraction than those in the RCT treatment, which is similar to the result obtained in Section 2.3.4. As expected, the estimates based on the mid-point imputation roughly lie between the estimates based on the left and right end point imputation. Also the left end point imputation generally gives the shortest survival estimate among the three single imputation approaches. It is interesting to see from Figure 2.3 that for both groups, the multiple imputation approach with three different values of M gives almost identical estimates, which are similar to those given by the mid-point imputation approach.

For the multiple imputation estimates given in Figure 2.3, the estimate given by the left end point imputation is used as the initial estimate. For convergence, we apply the criterion

$$\sum_j \mid \hat{S}^{(l)}(t_j) - \hat{S}^{(l-1)}(t_j) \mid \leq \epsilon ,$$

where the t_j's are the ordered, distinct time points of all left and right end points of observed intervals, and ϵ is a prespecified positive number, taken to be 0.0001 here.

For the treatment comparison, we assume that time to breast retraction follows model (1.4) with the Z_i's defined as in Section 2.3.4 and β representing the treatment difference. Table 2.2 presents the results given by the single and multiple imputation approaches, respectively. For the multiple imputation approach, for comparison, the results are obtained with M equal to 30, 50, or 100. All six analyses give similar results about the treatment comparison and suggest that the adjuvant chemotherapy significantly increases the risk of breast retraction. Although $\hat{\beta}$ changes as M increases from 30 to 100, the estimated treatment effect seems to become stable when M approaches 50.

For a second example, we discuss the interval-censored data given in Table 2.3 about time to drug resistance to zidovudine. The data are reproduced from Lindsey and Ryan (1998) and Richman et al. (1990) and consist of 31 AIDS patients enrolled in four clinical trials for the treatment of AIDS. Because the resistance assays were very expensive, few assessments were performed on each patient. Consequently, like the data in Table 2.1, this is a set of highly interval-censored observations, and there is a high proportion of right-censored observations.

Table 2.2. Estimated effects of adjuvant chemotherapy on time to breast retraction

Method	$\hat{\beta}$	SD($\hat{\beta}$)	p-value
Left end point imputation	0.9120	0.3483	0.009
Middle point imputation	0.9001	0.3454	0.009
Right end point imputation	0.7681	0.3486	0.028
Multiple imputation with $M = 30$	0.9557	0.3474	0.006
Multiple imputation with $M = 50$	0.9237	0.3470	0.008
Multiple imputation with $M = 100$	0.9041	0.3481	0.009

Table 2.3. Observed intervals $(L, R]$ for time to zidovudine resistance for 31 AIDS patients and values of four associated covariates

L	R	Stage	Dose	CD4$_1$	CD4$_2$	L	R	Stage	Dose	CD4$_1$	CD4$_2$
0	16	0	0	0	1	0	15	0	1	1	0
14	26	0	0	0	1	2	26	1	0	0	1
11	26	0	0	0	1	3	26	1	0	0	1
16	26	0	0	0	1	0	11	1	0	0	1
12	26	0	0	0	1	12	19	1	0	0	1
0	24	0	0	1	0	0	6	1	0	0	1
5	26	0	1	1	0	0	11	1	1	0	0
0	15	0	1	1	0	5	26	1	1	0	0
13	26	0	1	1	0	0	6	1	1	0	0
11	26	0	1	1	0	1	12	1	1	0	0
12	26	0	1	0	1	0	17	1	1	1	0
11	26	0	1	1	0	0	14	1	1	0	0
11	26	0	1	1	0	0	25	1	1	0	1
0	18	0	1	0	1	1	11	1	1	0	0
0	14	0	1	0	1	0	14	1	1	0	0
0	17	0	1	1	0						

The data include four covariates, the stage (0 or 1, earlier or later stage of the disease), dose (0 or 1, lower or higher dose of zidovudine), and ranges of CD4 counts at randomization given by CD4$_1$ and CD4$_2$. Here CD4$_1$ indicates (by 1) if CD4 is between 100 and 399, and CD4$_2$ is the indicator of CD4 $\geq$ 400. The analysis results obtained using the single and multiple imputation approaches are given in Table 2.4, with $M = 50$ for the multiple imputation approach. Unlike those obtained for the breast cancer data, the results here differ, although none of estimated covariate effects is significant except in one instance. Possible explanations are that the sample size is small, and the observed information is very limited due to interval-censoring.

Another reason for differing results could be that the covariates are correlated. For example, the last two covariates, indicators of CD4 counts, are correlated and both are also correlated with stage of disease. For this reason, we remove the last two covariates and reanalyze the data. The analysis results are presented in Table 2.5. It can be seen that the estimated effects are much closer to each other than those given in Table 2.4. Both mid-point imputation and multiple imputation approaches suggest that the patients in the later stage of the disease have significantly higher risk of developing resistance to zidovudine than those in the earlier stage of the disease.

Table 2.4. Estimated effects of four covariates on time to zidovudine resistance

Method	$\hat{\beta}$	SD($\hat{\beta}$)	p-value
Stage of disease			
Left end point imputation	0.8755	0.7687	0.255
Middle point imputation	1.0083	0.7280	0.166
Right end point imputation	0.0198	0.8279	0.981
Multiple imputation	1.2548	0.9709	0.196
Dose of zidovudine			
Left end point imputation	0.6394	0.9182	0.486
Middle point imputation	0.1232	0.8794	0.889
Right end point imputation	0.0215	0.7168	0.976
Multiple imputation	0.3492	1.0836	0.747
$100 \leq \text{CD4} \leq 399$			
Left end point imputation	0.1446	0.9451	0.878
Middle point imputation	-1.3721	1.0931	0.209
Right end point imputation	-2.1488	1.2300	0.763
Multiple imputation	-0.3662	1.2151	0.763
$\text{CD4} \geq 400$			
Left end point imputation	0.1459	1.0031	0.884
Middle point imputation	-1.6082	1.1169	0.150
Right end point imputation	-2.3043	1.0725	0.032
Multiple imputation	-0.6369	1.2417	0.608

Table 2.5. Estimated effects of two covariates on time to zidovudine resistance

Method	$\hat{\beta}$	SD($\hat{\beta}$)	p-value
Stage of disease			
Left end point imputation	0.7925	0.5821	0.173
Middle point imputation	1.3980	0.5906	0.019
Right end point imputation	0.7709	0.5948	0.195
Multiple imputation	1.4056	0.6762	0.038
Dose of zidovudine			
Left end point imputation	0.5672	0.6340	0.371
Middle point imputation	0.5333	0.6879	0.438
Right end point imputation	0.3133	0.6301	0.619
Multiple imputation	0.4572	0.6764	0.499

2.5 Bibliography, Discussion, and Remarks

The literature on parametric and imputation approaches for interval-censored failure time data is relatively limited although the idea behind them seems to be straightforward. Specifically, the authors who studied imputation methods for interval-censored data include Bebchuk and Betensky (2000), Betensky and Finkelstein (1999a), Dorey et al. (1993), Pan (2000a, b, 2001), and Satten et al. (1998). For parametric inference approaches, in addition to Lindsey

(1998) and Lindsey and Ryan (1998), others that investigated them include Boardman (1973), Bogaerts et al. (2002), Burridge (1981, 1982), Farrington (1996), Hazelrig et al. (1982), Kooperberg and Clarkson (1997), Marshall (1974), Odell et al. (1992), Samuelson and Kongerud (1994), Tikhov (2004), and Younes and Lachin (1997). In particular, among others, Farrington (1996), Kooperberg and Clarkson (1997), Lindsey and Ryan (1998), and Younes and Lachin (1997) proposed some so-called weakly parametric models. The weakly parametric model usually refers to a model that is parametric in theory but can provide good approximations to nonparametric models with the increasing of the dimension of space that they belong to.

Suppose that the half line $[0, \infty)$ is divided into J different intervals $\{I_j = [\tau_j, \tau_{j+1})\,;\, j = 1, ..., J\}$. Farrington (1996) and Lindsey and Ryan (1998) considered the piecewise exponential model that assumes that the hazard function has the form

$$\lambda(t) = \lambda_j \ , \ t \in [\tau_j, \tau_{j+1}) \ ,$$

$j = 1, ..., J$. Here the λ_j's are unknown parameters and $\tau_1 = 0 < \tau_2 < ... < \tau_J < \tau_{J+1} = \infty$. If there exists a vector of covariates, $\boldsymbol{Z}$, this model can be generalized to

$$\lambda(t\,;\, \boldsymbol{Z}) = \lambda_j \, \exp(\boldsymbol{Z}'\boldsymbol{\beta}) \ , \ t \in [\tau_j, \tau_{j+1})$$

or

$$\lambda(t\,;\, \boldsymbol{Z}) = \lambda_j + \boldsymbol{Z}'\boldsymbol{\beta} \ , \ t \in [\tau_j, \tau_{j+1})$$

following model (1.4) or (1.7), where $\boldsymbol{\beta}$ denotes regression parameters as before. Similar to the these models, Kooperberg and Clarkson (1997) suggested using

$$\log \lambda(t\,;\, \boldsymbol{Z}) = \sum_{j=1}^{J} \beta_j \, B_j(t \,|\, \boldsymbol{Z})$$

given $\boldsymbol{Z}$, and Younes and Lachin (1997) proposed a class of models in which the baseline hazard function can be expressed as

$$\lambda_0(t) = \sum_{j=1}^{J} e^{\beta_j} \, B_j(t) \ .$$

In these models, the β_j's are unknown parameters, the $B_j(t \,|\, \boldsymbol{Z})$'s are known base functions of some function space, and the $B_j(t)$'s are known B-spline functions.

The advantage of weakly parametric models is their flexibility. For example, when J in the models described above increases, they become close to nonparametric models in nature. At the same time, they still have the advantages of parametric models. The inference approach based on these models provides a bridge and a compromise between purely nonparametric or semiparametric approaches and parametric approaches.

Parametric and imputation methods are two different types of approaches, but for the analysis of interval-censored data, the idea behind them is the same: to change to a more familiar or simpler model. As mentioned before, a key advantage of parametric inference approaches is that their implementation is straightforward in principle and standard maximum likelihood theory generally applies. They provide attractive choices in particular if censored intervals are very wide and/or sample sizes are small, resulting in very limited information about survival variables of interest. A major disadvantage of these methods is that there often does not exist enough prior information or data to suggest or verify a parametric model. The main advantage of imputation approaches is that by using them, one can avoid complex interval censoring and also can make use of the existing inference procedures and statistical software for right-censored data. Some drawbacks include biased estimates and underestimation of the variability of point estimates.

As mentioned before, for all the methods discussed in this chapter and in this book except in Section 10.5, it is assumed that the censoring mechanism or variables are independent of the survival variables of interest. That is, one has independent interval censoring. For dependent or informative interval censoring, the imputation approach may not be appropriate because it could destroy the dependent structure. In contrast, one could use parametric models for both failure time variables and censoring variables although it may not be easy to determine appropriate models. More discussion on this topic is given in Section 10.5.

For the parametric inference approaches discussed in Section 2.3, some asymptotic approximations to the distributions of the maximum likelihood estimator and the score and likelihood ratio statistics are described without giving their derivations. To obtain these derivations, central limit theorems for sums of independent random variables are needed along with some regularity conditions. For instance, it is commonly required that the true values of the unknown parameters lie in the interior of the parameter space. For discussion on such theorems and conditions, see, for example, Feller (1971) and Shorack (2000). Assessment of the goodness-of-fit of data to a specified parametric model also is not discussed in this chapter. In other words, in the use of parametric approaches, methods are usually needed for the selection of an appropriate model or the best parametric model among several competitors. Some general goodness-of-fit approaches for this are discussed in Section 10.2.

This page intentionally blank

3

Nonparametric Maximum Likelihood Estimation

3.1 Introduction

Estimation of a survival function is perhaps the first and most commonly required task in the analysis of failure time data. There can be many reasons or purposes for such a task. For example, an estimated survival function can be used to assess the validity of an assumption about a particular parametric model for the underlying survival variable of interest. Also, one may need to estimate survival functions to estimate certain survival probabilities, to graphically compare several different treatments, or to predict survival probabilities for future patients. In the case where a parametric model can be reasonably assumed for the underlying survival function, the estimation problem is relatively easy, and the maximum likelihood approach discussed in Section 2.3 is commonly used for the problem. In this chapter, attention is focused on nonparametric estimation of survival functions along with estimation of hazard functions.

In the case of right-censored failure time data, the nonparametric maximum likelihood estimator (NPMLE) of a survival function is given by the Kaplan-Meier estimator (Kaplan and Meier, 1958; Kalbfleisch and Prentice, 2002). It is a product-limit estimator and has been extensively studied in the literature. Furthermore, its pointwise variance estimate is available and given by the well-known Greenwood's formula (Greenwood, 1926). For interval-censored failure time data, unlike parametric inference, nonparametric inference is much more complicated than that for right-censored data from both practical and theoretical points of views. In particular, the NPMLE of a survival function does not have a closed form in general and can only be determined using iterative algorithms.

Sometimes, one may also be interested in estimation of a hazard function, which could give more insight about the variable of interest than its survival function. For this, it is common and natural to apply weakly parametric approaches, or smoothing estimation techniques such as kernel and spline estimation. This is because parametric approaches usually involve model selection or checking, and purely nonparametric approaches often give estimates that are rough and not very helpful. This chapter discusses several smoothing estimation approaches with the focus on kernel estimation for interval-censored failure time data.

Section 3.2 deals with the NPMLE of a survival function based on case I interval-censored or current status data. For this special type of interval-censored data, a closed form is available for the NPMLE. Section 3.3 considers the NPMLE of a survival function based on general interval-censored failure time data and discusses characterization of the NPMLE that is useful for its determination. In Section 3.4, three algorithms for determination of the NPMLE for case II interval-censored data are described. They are the self-consistency algorithm, the iterative convex minorant (ICM) algorithm, and the EM iterative convex minorant (EM-ICM) algorithm. The smooth estimation of hazard functions is considered in Section 3.5. Asymptotic properties of the NPMLE are the topic of Section 3.6 with attention confined to discussion of the properties and the conditions required of the observation process, but not their derivations. Section 3.7 provides bibliographic notes about nonparametric maximum likelihood estimation for interval-censored data and discusses some related issues and topics that are not treated in this chapter.

3.2 NPMLE for Current Status Data

This section assumes that case I interval-censored, i.e., current status, failure time data are available from n independent subjects. Let the T_i's denote the survival times of interest with survival function $S(t)$ and assume that the observed data have the form

$$\{\,(C_i\,,\,\delta_i)\;\; i\,=\,1,...,n\,\}\,,$$

where C_i denotes the observation time for subject i independent of T_i and $\delta_i\,=\,I(T_i\,\leq\,C_i)$. Then the likelihood function has the form

$$L_S(\,S(t)\,)\,=\,\prod_{i=1}^{n}\,[\,S(C_i)]^{1-\delta_i}\,[\,1\,-\,S(C_i)\,]^{\delta_i}\;.$$

Let $\{\,s_j\,\}_{j=0}^{m}$ denote the unique ordered elements of $\{\,0, C_i\,;\,i\,=\,1,...,n\,\}$. Define $r_j\,=\,\sum_{i=1}^{n}\,\delta_i\,I(C_i\,=\,s_j)$, the number of subjects who are observed at s_j and found to have failed, and $n_j\,=\,\sum_{i=1}^{n}\,I(C_i\,=\,s_j)$, the number of subjects who are observed at s_j, $j\,=\,1,...,m$. Then the likelihood function $L_S(\,S(t)\,)$ can be rewritten as

$$L_S(S(t))\,=\,\prod_{j=1}^{m}\,[S(s_j)]^{n_j-r_j}\,[1-S(s_j)]^{r_j}\,=\,\prod_{j=1}^{m}\,[F(s_j)]^{r_j}\,[1-F(s_j)]^{n_j-r_j}\;,$$

which is proportional to the likelihood arising in an m-sample binomial setting, where $F(t) = 1 - S(t)$.

It is apparent that the likelihood function L_S depends on S or F only through its values at the s_j's. That is, one can estimate S or F only at these s_j's. By noting the constraint $F(s_1) \leq ... \leq F(s_m)$, one can show that the maximization of $L_S(S(t))$ with respect to $\{ F(s_j) \}_{j=1}^m$ is equivalent to minimizing

$$\sum_{j=1}^{m} n_j \left[\frac{r_j}{n_j} - F(s_j) \right]^2$$

subject to $F(s_1) \leq ... \leq F(s_m)$ (Robertson et al., 1988). The set of values of $\{ F(s_j) \}_{j=1}^m$ that minimize this summation is commonly referred to as the isotonic regression of $\{r_1/n_1, ..., r_m/n_m\}$ with weights $\{n_1, ..., n_m\}$ (Barlow et al., 1972; Robertson et al., 1988). Using the max-min formula for isotonic regression, the NPMLE of F at time s_j has the value

$$\hat{F}(s_j) = \max_{u \leq j} \min_{v \geq j} \frac{\sum_{l=u}^{v} d_j}{\sum_{l=u}^{v} n_j} ,$$

giving the value of the NPMLE of S at s_j to be $1 - \hat{F}(s_j)$. That is, the NPMLE of S has a closed form. To computer $\hat{F}(s_j)$, one also can use other algorithms for the isotonic regression such as the pool adjacent violators algorithm.

For illustration, we apply the formula given above for current status data to the lung tumor data discussed in Section 1.2.1. The data consist of 144 male mice from two treatment groups, conventional environment and germ-free environment. For each animal, the death time, which serves as observation

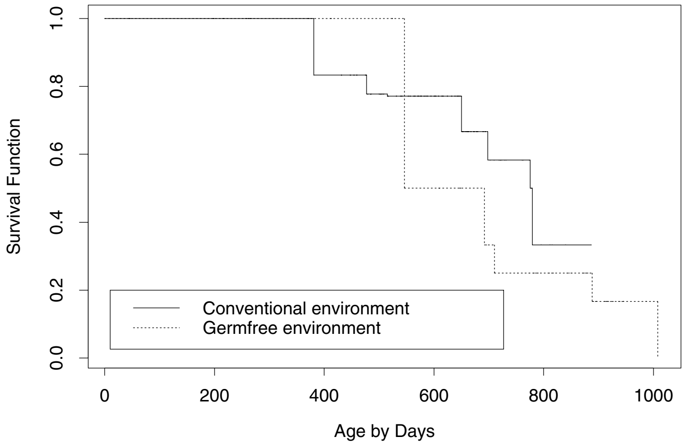

Fig. 3.1. Estimates of survival functions of time to lung tumor onset.

time, and the presence and absence of lung tumors are observed. Assume that the death time is independent of time to lung tumor onset. Figure 3.1 presents the NPMLE of the survival functions of the time to lung tumor onset for the two groups separately. The estimates suggest that a lung tumor seems to occur earlier for the mice in the conventional environment than those in the germ-free environment. However, overall the mice in the germ-free environment seem to have a higher risk to develop a lung tumor than those in the conventional environment.

For general or case II interval-censored data, there is no closed form available for the NPMLE of S. In this situation, an iterative algorithm given in Section 3.4 has to be applied.

3.3 Characterization of NPMLE for Case II Interval-censored Data

Consider a failure time study that consists of n independent subjects from a homogeneous population with survival function $S(t)$. Let T_i denote the survival time of interest for subject i, $i = 1, ..., n$. Suppose that interval-censored data on the T_i's are observed and given by

$$\boldsymbol{O} = \{ (L_i, R_i] ; i = 1, ..., n \} ,$$

where $(L_i, R_i]$ denotes the interval to which T_i is observed to belong. Also suppose that the goal is to derive the NPMLE of $S(t)$.

Let $\{ s_j \}_{j=0}^m$ denote the unique ordered elements of $\{ 0, L_i, R_i ; i = 1, ..., n \}$ as in the previous section. Define $\alpha_{ij} = I(s_j \in (L_i, R_i])$ and $p_j = S(s_{j-1}) - S(s_j)$, $i = 1, ..., n$, $j = 1, ..., m$. Then the likelihood function is

$$L_S(\boldsymbol{p}) = \prod_{i=1}^n [S(L_i) - S(R_i)] = \prod_{i=1}^n \sum_{j=1}^m \alpha_{ij} \, p_j , \tag{3.1}$$

where $\boldsymbol{p} = (p_1, ..., p_m)'$.

As with current status data, it is easy to show that the likelihood function L_S depends on S only through the values $\{ S(s_j) \}_{j=1}^m$ and not how S changes between the s_j's. In other words, the NPMLE of S can be uniquely determined only up to its values at these s_j's, and its determination is equivalent to maximizing $L_S(\boldsymbol{p})$ with respect to $\boldsymbol{p}$ subject to the constraints

$$\sum_{j=1}^m p_j = 1 ,$$

$$p_j \geq 0 \ \ (j = 1, ..., m)$$

(Gentleman and Geyer, 1994; Turnbull, 1976).

In the following, as usual, it is assumed that the NPMLE of S or of $F(t) = 1 - S(t)$ corresponds with a discrete distribution with jumps only at $\{ s_j \}_{j=1}^{m}$ unless otherwise specified. Then one determines the NPMLE by maximizing the likelihood function given in (3.1) over all discrete survival or distribution functions that are constant between the points $s_0 < s_1 < ... < s_m$. In some situations, it may be of interest to maximize the likelihood over a different set of survival or distribution functions. For example, sometimes it may be reasonable to assume that F is smooth and thus one wants the NPMLE to be a smooth function. More comments on this are given at the end of the next section.

For determination of the NPMLE of S or F, one needs to choose an algorithm, such as the one discussed in the next section. In addition, one may need to consider some related issues. For instance, if $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_m)'$ is the NPMLE, then some elements of $\hat{\boldsymbol{p}}$ could be zero and it could help greatly to know these zero components before running a determination program. For this, one can use the fact that each $\hat{p}_j$ can be nonzero only if $s_{j-1} = L_i$ for some i and $s_j = R_k$ for some possibly different k, $i, k = 1, ..., n$ (Peto, 1973; Turnbull, 1976). Thus one only needs to focus on the $\hat{p}_j$'s that satisfy this condition, but of course, some of them could still be zero. The use of this fact could considerably reduce the number of time points, m, that need to be considered as well as the computational effort. Furthermore, one may use the Lagrange multiplier criterion described below.

In terms of determining the possible nonzero $\hat{p}_j$'s, an alternative approach is to find all time points s_j's or a set of disjoint intervals that constitutes the possible support of the NPMLE of S or F. From the discussion above, these intervals are the ones whose left and right end points are given by some of the L_i's and R_i's, respectively, and that contain no other L_i's and R_i's except at their end points. This is often referred to as Turnbull's approach, and from this point of view, the determination of the NPMLE consists of two steps. The first one is to determine the possible support points or intervals and the second step is to maximize $L_S(\boldsymbol{p})$.

To determine if a candidate estimate $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_m)'$ of $\boldsymbol{p}$ is the NPMLE, define

$$d_j(\boldsymbol{p}) = \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{l=1}^{m} \alpha_{il} \, p_l} \, , \tag{3.2}$$

$j = 1, ..., m$. The Lagrange multiplier criterion, derived from graph theory, says that $\hat{\boldsymbol{p}}$ is the NPMLE if $d_j(\hat{\boldsymbol{p}}) = n$ for all $j = 1, ..., m$ (Gentleman and Geyer, 1994). Furthermore, using the general mixture maximum likelihood theorem, one can in fact show that $\hat{\boldsymbol{p}}$ is the NPMLE if and only if $d_j(\hat{\boldsymbol{p}}) \leq n$ for all j (Böhning et al., 1996).

To illustrate the Lagrange multiplier criterion, consider the data set

$$\{ (0, 1], (1, 3], (1, 3], (0, 2], (0, 2], (2, 3] \}$$

from Gentleman and Geyer (1994). It consists of six observed intervals and is also discussed by Böhning et al. (1996). For this data, one has $\{ s_j \}_{j=0}^{3} = \{ 0, 1, 2, 3\}$,

$$A = (\alpha_{ij}) = \begin{pmatrix} 1\,0\,0 \\ 0\,1\,1 \\ 0\,1\,1 \\ 1\,1\,0 \\ 1\,1\,0 \\ 0\,0\,1 \end{pmatrix}$$

and

$$L_S(\boldsymbol{p}) = \prod_{i=1}^{6} (\alpha_{i1}\,p_1 + \alpha_{i2}\,p_2 + \alpha_{i3}\,p_3) \ .$$

To determine the NPMLE, one needs to maximize the likelihood function L_S with respect to $\boldsymbol{p} = (p_1, p_2, p_3)'$ subject to $\sum_{j=1}^{3} p_j = 1$ and $p_1, p_2, p_3 \geq 0$. Suppose that one is given two estimators: $\hat{\boldsymbol{p}}_1 = (1/2, 0, 1/2)'$ and $\hat{\boldsymbol{p}}_2 = (1/3, 1/3, 1/3)'$. To check if one is the NPMLE, note that

$$d_1 = \frac{1}{p_1} + \frac{2}{p_1 + p_2}\,,\ d_2 = \frac{2}{p_2 + p_3} + \frac{2}{p_1 + p_2}\,,\ d_3 = \frac{2}{p_2 + p_3} + \frac{1}{p_3}\,.$$

Thus it is clear that $\hat{\boldsymbol{p}}_1$ is not the NPMLE because $d_2 = 8 > n = 6$, while $\hat{\boldsymbol{p}}_2$ satisfies the Lagrange multiplier criterion and thus is the NPMLE because $d_1 = d_2 = d_3 = 6 = n$.

The uniqueness of the NPMLE is another issue that often needs to be checked. For this, note that log $L_S(\boldsymbol{p})$ is a concave function but may not be strictly concave. Define $A = (\,\alpha_{ij}\,)$, the $n \times m$ matrix. Then the NPMLE is unique if the rank of A is equal to m. This guarantees that the log likelihood function, log $L_S(\boldsymbol{p})$, is strictly concave and thus has a unique maximum. In practice, the rank of A may often be less than m. For this, let A^* denote the submatrix of A that consists of all the columns of A such that either the corresponding $p_j > 0$ or $d_j(\boldsymbol{p}) = n$ if the corresponding $p_j = 0$. Then a sufficient condition for the uniqueness of the NPMLE, given in Gentleman and Geyer (1994), is that the rank of A^* is equal to its number of columns.

3.4 Algorithms for Case II Interval-censored Data

This section first presents three algorithms that can be applied to determine the NPMLE of S, F or $\boldsymbol{p}$. The first one is the self-consistency algorithm that was developed by Turnbull (1976) and can be regarded as an application of the EM algorithm (Dempster et al., 1977). The second algorithm is the ICM algorithm, first introduced by Groeneboom and Wellner (1992) and later modified by Jongbloed (1998). It transforms maximization of the likelihood function (3.1) to maximization of a quadratic function using isotonic regression theory. The third algorithm is a hybrid algorithm proposed by Wellner and Zhan

(1997), which is referred to as the EM-ICM algorithm in the following. It basically combines the self-consistency algorithm and the ICM algorithm. Finally, two illustrative examples are provided and followed by some general discussion about the estimation of S or F.

3.4.1 The Self-consistency Algorithm

A self-consistent estimate usually refers to an estimate that can be characterized by a self-consistency equation and is the limit of iterates obtained from that equation (Efron, 1967). Consequently, the estimate can be determined iteratively. To derive the self-consistency equation for interval-censored data, one may use a direct and intuitive approach based on empirical estimates. A more general approach is to treat interval-censored data as incomplete data and then to apply the EM algorithm. In the following, the general approach is used and is followed by some comments on the direct approach.

Define

$$\boldsymbol{C_p} = \left\{ \boldsymbol{p} \in [0,1]^m \, ; \, \sum_{j=1}^{m} p_j = 1 \, , \, p_j \geq 0 \right\} ,$$

a subspace of $\mathcal{R}^m$. One determines the NPMLE of $\boldsymbol{p}$ by maximizing the likelihood function $L_S(\boldsymbol{p})$ given in (3.1) over the region $\boldsymbol{C_p}$. To apply the EM algorithm, suppose that exact failure time data $\{ T_i \}_{i=1}^n$ are available. The log-likelihood function for these complete data is

$$l_S(\boldsymbol{p} \, ; T_1, ..., T_n) = \log \left[\prod_{i=1}^{n} dF(T_i) \right] = \sum_{j=1}^{m} d_j^* \log p_j \, ,$$

where $d_j^* = \sum_{i=1}^n I(T_i = s_j)$, $j = 1, ..., m$. Let $\hat{\boldsymbol{p}}^{\,c}$ denote the current estimate of $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}^{\,u}$ the updated estimate of $\boldsymbol{p}$. In the E-step, one needs to calculate the conditional expectation of $l_S(\boldsymbol{p} \, ; T_i's)$ given $\hat{\boldsymbol{p}}^{\,c}$ and the observed data $\boldsymbol{O}$, which has the form

$$E \left[l_S(\boldsymbol{p} \, ; T_i's) | \hat{\boldsymbol{p}}^{\,c}, \boldsymbol{O} \right] = \sum_{j=1}^{m} \log(p_j) \, E(d_j^* | \hat{\boldsymbol{p}}^{\,c}, \boldsymbol{O}) = \sum_{j=1}^{m} d_j(\hat{\boldsymbol{p}}^{\,c}) \, p_j^c \, \log(p_j) \, ,$$

where $d_j(\boldsymbol{p})$ is defined in (3.2).

For the M-step in the EM algorithm, one needs to maximize the conditional expectation given above over the region $\boldsymbol{C_p}$. Using the Lagrange approach, one maximizes

$$\sum_{j=1}^{m} d_j(\hat{\boldsymbol{p}}^{\,c}) \, p_j^c \, \log(p_j) + \lambda \left(1 - \sum_{j=1}^{m} p_j \right)$$

over $\boldsymbol{p}$ and λ. Differentiating the function above with respect to p_j and setting the derivatives to 0, one gets $p_j = d_j(\hat{\boldsymbol{p}}^{\,c}) \, p_j^c / \lambda$. It then follows from $\sum_{j=1}^m p_j = 1$ that $\lambda = n$ and

$$\hat{p}_j^u = \frac{d_j(\hat{\boldsymbol{p}}^c)\, p_j^c}{n} = \frac{1}{n} E\left[\sum_{i=1}^n I(T_i = s_j) \,|\, \hat{\boldsymbol{p}}^c, \boldsymbol{O} \right] , \tag{3.3}$$

which suggests the following self-consistency algorithm for the NPMLE.
Step 1. Choose an initial estimate $\hat{\boldsymbol{p}}^0$ of $\boldsymbol{p}$.
Step 2. At the lth iteration, define the updated estimate, denoted by $\hat{\boldsymbol{p}}^{(l)} = (\hat{p}_1^{(l)}, ..., \hat{p}_m^{(l)})'$, of $\boldsymbol{p}$ as

$$\hat{p}_j^{(l)} = \frac{d_j(\hat{\boldsymbol{p}}^{(l-1)})\, p_j^{(l-1)}}{n} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha_{ij}\, \hat{p}_j^{(l-1)}}{\sum_{k=1}^m \alpha_{ik}\, \hat{p}_k^{(l-1)}} ,$$

$j = 1, ..., m$.
Step 3. Repeat step 2 until the desired convergence occurs.

The estimate $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_m)'$ given by this algorithm is the solution to the self-consistency equation

$$\hat{p}_j = \frac{1}{n} E\left[\sum_{i=1}^n I(T_i = s_j) \,|\, \hat{\boldsymbol{p}}, \boldsymbol{O} \right] \tag{3.4}$$

and is a self-consistent estimate. Also, it can be shown that the likelihood function increases after each iteration. Although the estimate $\hat{p}$ may not be the NPMLE of $\boldsymbol{p}$, it can be checked using the criterion discussed in Section 3.3.

A direct approach that also gives the self-consistency algorithm and equation (3.4) is to look at the second term in equation (3.3). Given $\boldsymbol{p} = \hat{\boldsymbol{p}}^c$, the quantity represents n^{-1} times the estimated or expected number of subjects whose survival times are equal to s_j, which naturally yields the algorithm based on empirical estimates. With $\hat{F}(t) = \sum_{s_j \leq t} \hat{p}_j$, equation (3.4) gives

$$\hat{F}(t) = \frac{1}{n} E\left[\sum_{i=1}^n I(T_i \leq t) \,|\, \hat{F}, \boldsymbol{O} \right] . \tag{3.5}$$

3.4.2 The Iterative Convex Minorant Algorithm

To describe the ICM algorithm, define

$$C_{\boldsymbol{x}} = \left\{ \boldsymbol{x} = (x_1, ..., x_{m-1})' \in \mathcal{R}^{m-1} ; 0 \leq x_1 \leq ... \leq x_{m-1} \leq 1 \right\} ,$$

a subspace of $\mathcal{R}^{m-1}$, and let $\beta_j = F(s_j)$, $j = 1, ..., m$. With $\beta_0 = 0$, $\beta_m = 1$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_{m-1})'$, the likelihood function in (3.1) can be rewritten as

$$L_S(\boldsymbol{\beta}) = \prod_{i=1}^n \sum_{j=1}^m \alpha_{ij} \left(\beta_j - \beta_{j-1} \right) , \tag{3.6}$$

and the NPMLE is obtained by maximizing $L_S(\boldsymbol{\beta})$ over $C_{\boldsymbol{x}}$.

The ICM algorithm is based on the following two facts. First, suppose that g is a differentiable, concave mapping from $\mathcal{R}^{m-1}$ to $\mathcal{R}$ and C is a convex cone in $\mathcal{R}^{m-1}$. Also suppose that $g(\boldsymbol{x})$ achieves its maximum over region C at $\hat{\boldsymbol{x}}$. Let $\boldsymbol{W}$ be a positive definite $(m-1) \times (m-1)$ matrix and $\boldsymbol{y}$ a fixed point in $\mathcal{R}^{m-1}$. Define

$$g^*(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{W}) \;=\; -\,\frac{1}{2}\,(\boldsymbol{x} \,-\, \boldsymbol{y})'\,\boldsymbol{W}\,(\boldsymbol{x} \,-\, \boldsymbol{y})$$

for $\boldsymbol{x} \in \mathcal{R}^{m-1}$ and suppose that $\hat{\boldsymbol{x}}^* \in C$ maximizes $g^*(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{W})$ over C. Of course, $\hat{\boldsymbol{x}}^*$ may depend on $\boldsymbol{y}$ and $\boldsymbol{W}$. With $\boldsymbol{y} \,=\, \hat{\boldsymbol{x}} \,+\, \boldsymbol{W}^{-1} \,\bigtriangledown g(\hat{\boldsymbol{x}})$, where $\bigtriangledown g(\boldsymbol{x})$ denotes the vector of derivatives of g at $\boldsymbol{x}$, $\hat{\boldsymbol{x}}^*$ maximizes $g^*(\boldsymbol{x} \,|\, \boldsymbol{y}, \boldsymbol{W})$ over C if and only if $\hat{\boldsymbol{x}}^* \,=\, \hat{\boldsymbol{x}}$ (Groeneboom and Wellner, 1992).

The second fact concerns maximization of a quadratic function over region $C_{\boldsymbol{x}}$. Let $\hat{\boldsymbol{x}}^* \,=\, (\hat{x}_1^*, ..., \hat{x}_{m-1}^*)'$ be defined as above with $C \,=\, C_{\boldsymbol{x}}$ and $\boldsymbol{W} \,=\, diag(w_j)$ be a positive definite diagonal matrix. Define $P_0 \,=\, (0, 0)$ and

$$P_u \;=\; \left(\sum_{i=1}^{u} w_i \,,\, \sum_{i=1}^{u} w_i y_i \right) \;,\, 1 \,\leq\, u \,\leq\, m-1 \,,$$

points in $\mathcal{R}^2$ for some fixed $\boldsymbol{y} \,=\, (y_1, ..., y_{m-1})' \,\in\, \mathcal{R}^{m-1}$. The set of points $\{\, P_u\,;\, u \,=\, 0, ..., m \,-\, 1 \,\}$ is commonly referred to as a cumulative sum diagram because the coordinates of P_u are cumulative sums of the vectors $(w_1, ..., w_{m-1})'$ and $(w_1 y_1, ..., w_{m-1} y_{m-1})'$. Then $\hat{x}_i^*$ is given by the left derivative of the convex minorant of, i.e., the largest convex function below the cumulative sum diagram $\{\, P_u\,;\, u = 0, ..., m-1 \,\}$ evaluated at P_j.

The first fact suggests that if $\hat{\boldsymbol{x}}$ is known, the maximization of a general function $g(\boldsymbol{x})$ is equivalent to the maximization of the quadratic function $g^*(\boldsymbol{x})$, which is usually relatively easy. The second fact gives the maximizing point $\hat{\boldsymbol{x}}^*$ for a special quadratic function. Of course, $\hat{\boldsymbol{x}}$ is unknown and the first fact does not give a direct maximization procedure, but it can be used in an iterative fashion. The two facts together motivate the ICM algorithm described below.

Step 1. Choose an initial estimate $\hat{\boldsymbol{\beta}}^0$ of $\boldsymbol{\beta}$.

Step 2. At the lth iteration, define the updated estimate denoted by $\hat{\boldsymbol{\beta}}^{(l)} \,=\, (\hat{\beta}_1^{(l)}, ..., \hat{\beta}_{m-1}^{(l)})'$ of $\boldsymbol{\beta}$ as the $\hat{\boldsymbol{x}}^*$ that maximizes $g^*(\boldsymbol{x} \,|\, \boldsymbol{y}, W(\hat{\boldsymbol{\beta}}^{(l-1)}))$ with

$$\boldsymbol{y} \;=\; \hat{\boldsymbol{\beta}}^{(l-1)} \;-\; W^{-1}(\hat{\boldsymbol{\beta}}^{(l-1)}) \bigtriangledown l_S(\hat{\boldsymbol{\beta}}^{(l-1)})$$

and $\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(l-1)})$ being a positive definite diagonal matrix that may depend on $\hat{\boldsymbol{\beta}}^{(l-1)}$, where $l_S(\boldsymbol{\beta})$ is the log likelihood function from (3.6). A choice for $\boldsymbol{W}$ is discussed later. In other words, $\hat{\boldsymbol{\beta}}^{(l)}$ is taken to be the derivative of the convex minorant of the cumulative sum diagram $\{P_u\,;\, u = 0, ..., m-1\}$ given by $P_0 \,=\, (0, 0)$ and

$$P_u = \left(\sum_{i=1}^{u} w_i^{(l-1)} , \sum_{i=1}^{u} \left(w_i^{(l-1)} \hat{\beta}_i^{(l-1)} - \frac{\partial}{\partial \beta_i} l_S(\hat{\boldsymbol{\beta}}^{(l-1)}) \right) \right)$$

for $1 \leq u \leq m-1$, where $w_i^{(l-1)}$ is the ith diagonal element of $W(\hat{\boldsymbol{\beta}}^{(l-1)})$.
Step 3. Return to step 2 until the desired convergence occurs.

Jongbloed (1998) shows that the ICM algorithm may not increase the log likelihood function after each iteration and converge globally, and suggests to add a line search into the algorithm based on the following fact to achieve global convergence. Let g, g^*, C and $\hat{\boldsymbol{x}}$ be as given in the first fact. For given $\boldsymbol{x}$ and a positive definite diagonal matrix $\boldsymbol{W}(\boldsymbol{x})$ that may depend on $\boldsymbol{x}$, also let $A(\boldsymbol{x})$ be vector $\boldsymbol{z}$ at which $g^*(\boldsymbol{z} \,|\, \boldsymbol{y}, W)$ achieves its maximum with

$$\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{W}^{-1}(\boldsymbol{x}) \bigtriangledown l_S(\boldsymbol{x}) \,.$$

Then for $\boldsymbol{x} \neq \hat{\boldsymbol{x}}$ and all sufficiently small $\lambda > 0$,

$$g\left(\boldsymbol{x} + \lambda \left(A(\boldsymbol{x}) - \boldsymbol{x}\right)\right) > g(\boldsymbol{x}) \,.$$

This suggests that a line search, given by step 2.1 below, can be incorporated into the ICM algorithm to guarantee that the log likelihood function increases and thus the global convergence of the algorithm.

Let $0 < \epsilon < 0.5$ be a fixed number controlling the line search process. The following step can be added between steps 2 and 3 of the ICM algorithm described above.
Step 2.1. If

$$l_S(\hat{\boldsymbol{\beta}}^{(l)}) > l_S(\hat{\boldsymbol{\beta}}^{(l-1)}) + (1-\epsilon) \, [\bigtriangledown l_S(\hat{\boldsymbol{\beta}}^{(l-1)})]' \, (\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}) \,,$$

then move to step 3. Otherwise, find a point $\boldsymbol{z}$ such as

$$\boldsymbol{z} = \hat{\beta}^{(l-1)} + \lambda \, (\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)})$$

for $0 \leq \lambda \leq 1$ that satisfies

$$\epsilon \, [\bigtriangledown l_S(\hat{\boldsymbol{\beta}}^{(l-1)})]' \, (\boldsymbol{z} - \hat{\boldsymbol{\beta}}^{(l-1)}) \leq l_S(\boldsymbol{z}) - l_S(\hat{\boldsymbol{\beta}}^{(l-1)})$$
$$\leq (1 - \epsilon) \, [\bigtriangledown l_S(\hat{\boldsymbol{\beta}}^{(l-1)})]' \, (\boldsymbol{z} - \hat{\boldsymbol{\beta}}^{(l-1)}) \,.$$

Let $\hat{\boldsymbol{\beta}}$ denote the estimate given by the ICM algorithm. Then the NPMLE of F and $\boldsymbol{p}$ are given by $\hat{F}(t) = \hat{\beta}_j$ if $s_j \leq t < s_{j+1}$ for $j = 0, ..., m-1$ and $\hat{p}_j = \hat{\beta}_j - \hat{\beta}_{j-1}$ for $j = 1, ..., m$, respectively. In the ICM algorithm, a natural choice for $W(\boldsymbol{\beta})$ is to take

$$w_j = w_j(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \beta_j^2} \, l_S(\boldsymbol{\beta}) \,,$$

assuming that it exists, $j = 1, ..., m-1$. Jongbloed (1998) studied this choice and others using simulation and suggested that it is better than the others and, in particular, the line search or step 2.1 seems to be used less under it.

3.4.3 The EM Iterative Convex Minorant Algorithm

This subsection introduces a third algorithm, the EM-ICM algorithm, that can be used to determine the NPMLE of F. It is a hybrid algorithm and simply combines the self-consistency and ICM algorithms together. Specifically,

Step 1. Choose an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$ or $\hat{\boldsymbol{p}}^0$ of $\boldsymbol{\beta}$ or $\boldsymbol{p}$.
Step 2. Apply steps 2 and 2.1 of the ICM algorithm to the current estimate to obtain an updated estimate.
Step 3. Apply step 2 of the self-consistency algorithm to the updated estimate given in step 2 above to obtain a new updated estimate. That is, apply step 2 of the self-consistency with $\hat{\boldsymbol{p}}^{(l-1)}$ being the updated estimate given in step 2 above.
Step 4. Go back to step 2 until the desired convergence occurs.

Using a general theorem about the global convergence of composite mappings, Wellner and Zhan (1997) show that the EM-ICM algorithm converges to the NPMLE if it exists and is unique and the log likelihood function is continuously differentiable. They also show using simulations that in applying the EM-ICM algorithm, one can omit the line search and still achieve the desired convergence.

To apply the algorithms described above, one needs to select a convergence criterion. A simple and natural one is to base the convergence on the closeness of consecutive estimates of F or $\boldsymbol{\beta}$, which can be measured by, for example,

$$\sum_{j=1}^{m-1} \mid \hat{\beta}_j^{(l)} - \hat{\beta}_j^{(l-1)} \mid < \epsilon \tag{3.7}$$

as in Section 2.3.4 or

$$\max_{1 \leq j \leq m-1} \mid \hat{\beta}_j^{(l)} - \hat{\beta}_j^{(l-1)} \mid < \epsilon \,,$$

where ϵ is a fixed, positive constant. Another criterion that is commonly used in maximizing a likelihood function is to base convergence on the change of the log likelihood function. In this case, the iterations stop if

$$\mid l_S(\hat{\boldsymbol{\beta}}^{(l)}) - l_S(\hat{\boldsymbol{\beta}}^{(l-1)}) \mid < \epsilon \,.$$

The two criteria given above can be easily implemented, but they cannot tell if the estimate given at convergence is a local or global maximizer. If there exist local maximizers or in general, one may prefer the criterion based on the Fenchel optimality conditions (Robertson et al., 1988). Under this criterion, one stops the iteration and accept $\hat{\boldsymbol{\beta}}^{(l)} = (\hat{\beta}_1^{(l)}, ..., \hat{\beta}_{m-1}^{(l)})'$ as the NPMLE of F if

$$\mid \sum_{j=1}^{m-1} \hat{\beta}_j^{(l)} \, \frac{\partial}{\partial \beta_j} \, l_S(\hat{\boldsymbol{\beta}}^{(l)}) \mid < \epsilon$$

and

$$\max\left\{\sum_{u=j}^{m-1}\frac{\partial}{\partial\beta_u}\,l_S(\hat{\boldsymbol{\beta}}^{(l)})\;;\;j=1,...,m-1\right\}<\epsilon\,.$$

3.4.4 Examples and Discussion

For illustration and comparison purposes, this subsection first applies the three algorithms described in the previous subsections to two examples and then provides some general discussion. First consider the interval-censored breast cancer data discussed in Sections 1.2.2, 2.3.4, and 2.4.3, and in particular, estimation of the survival function of the time to breast retraction. Figure 3.2 presents six NPMLEs of the two survival functions corresponding with the two treatments, RT and RCT, obtained separately using the algorithms. As expected, the NPMLEs of each survival function given by the three algorithms at their convergences are almost identical. The figure suggests that as seen before, the patients in the RCT group seem to develop breast retraction earlier than those in the RT group.

For the estimates in Figure 3.2, Splus 2000 PR2 on Windows XP is used for programming with the uniform distribution as the initial estimate. The convergence criterion used for all algorithms is criterion (3.7) with $\epsilon = 10^{-8}$. The numbers of iterations required by the self-consistency, modified ICM and EM-ICM algorithms are, respectively, 416, 38, and 4 for the RT group, and 505, 7, and 3 for the RCT group. That is, as expected, both the modified ICM

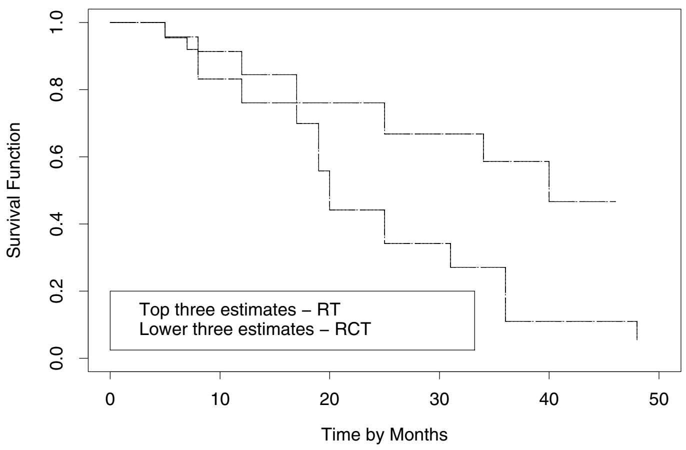

Fig. 3.2. NPMLE of survival functions of time to breast retraction.

and EM-ICM algorithms require fewer iterations than the self-consistency algorithm. In terms of computing time, however, the difference is relatively small as the CPU times required by the three algorithms are 0.43, 0.90, and 0.10 seconds for the RT group, and 0.55, 0.20, and 0.08 seconds for the RCT group, respectively. In particular, for the RT group, the self-consistency algorithm is faster than the modified ICM algorithm. This is because the self-consistency algorithm requires much less effort inside the iteration than the other two algorithms. We also used the Kaplan-Meier estimates based on imputed right-censored data as initial estimates and the numbers of iterations required for all the algorithms are different from, but similar to those given above.

For the second example, consider data set II of Appendix A. The data arise from a 16-center prospective study, described in detail by Goedert et al. (1989) and studied by Kroner et al. (1994) among others, to investigate HIV-1 infection risk among people with hemophilia. These patients were at risk of HIV-1 infection because they received for their treatments blood products such as factor VIII and factor IX concentrate made from the plasma of thousands of donors. In the study, for patients' HIV-1 infection times, only interval-censored data are available, and the patients are placed into different groups according to the average annual dose of the blood products they received. The data set II of Appendix A gives observations on 368 patients from five centers where patients were enrolled into the study without regard to their HIV-1 antibody status and who received no or low dose (between 1 and 20,000 U) factor VIII concentrate. The numbers of patients in the two groups are 236 and 132, respectively. For the data, the time unit is quarters and observation 0 means January 1, 1978, the start of the epidemic and the time at which all patients are considered to be negative.

For estimation of the survival function of the time to HIV-1 infection, Figure 3.3 displays six NPMLEs of the two survival functions for the patients in the two dose groups given by the three algorithms. As in Figure 3.2, the three NPMLEs of each survival function are almost identical. The figure indicates that the patients receiving low dose factor VIII concentrate seem to have significantly higher risk of being infected by HIV-1 than those receiving no factor VIII concentrate. For the estimates given here, the same programs, initial estimates and convergence criterion as those in Figure 3.2 are used. For the self-consistency, modified ICM and EM-ICM algorithms, the numbers of iterations required are, respectively, 354, 291, and 12 for the no factor VIII concentrate group and 2610, 6, and 4 for the low dose factor VIII concentrate group. The corresponding CPU times are, respectively, 0.56, 6.30, and 0.30 seconds for the former group, and 3.30, 0.17, and 0.11 seconds for the latter group.

It is interesting to note that compared with the first example, the algorithms seem to behave differently in terms of the number of iterations. For the low dose group, note that its sample size is larger than that of the data in the previous example, which naturally results in the large number of iterations needed by the self-consistency algorithm. In contrast, in this case, the

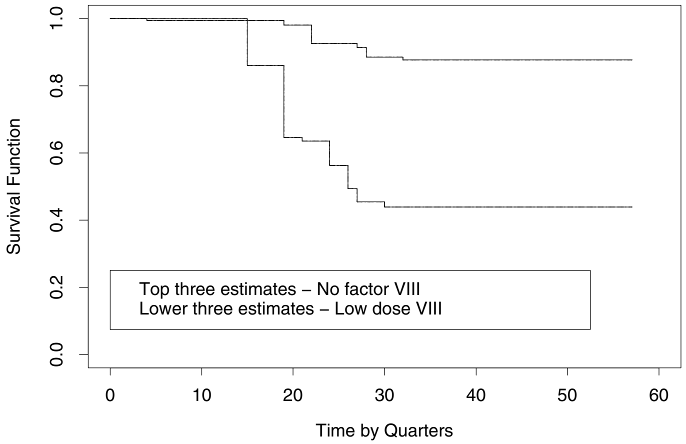

Fig. 3.3. NPMLE of survival functions of time to HIV-1 infection.

increase in sample size did not seem to affect the other two algorithms much. On the other hand, although the no factor VIII concentrate group involves many more patients, the numbers of iterations and the CPU times tell a different story. A close look at the data seems to indicate that this is because there exists a large proportion of right-censored observations for the group.

Among the self-consistency, ICM, and EM-ICM algorithms, the first algorithm is the simplest and the most natural one. Compared with the other two algorithms, it can be easily understood and implemented. Although it is slower than the other two, this usually does not produce a serious problem unless one faces a large data set or has to use it a number of times. One such situation occurs when one uses a bootstrap procedure to obtain confidence bands. In other words, for data sets with small or moderate sample sizes, although the self-consistency algorithm needs more iterations, it may still be a good choice given its simplicity and the actual computing time needed. The same is true for data with large proportions of right-censored observations. The drawback of the algorithm is that there is no guarantee that the resulting estimate is the NPMLE although it is definitely a self-consistent estimate.

The biggest advantage of the modified ICM and EM-ICM algorithms over the self-consistency algorithm is their rapid convergence in terms of the number of iterations and computing time required. More specifically, the EM-ICM algorithm is the fastest one among them. Both Jongbloed (1998) and Wellner and Zhan (1997) demonstrated these facts using simulation studies. Another advantage of these two over the self-consistency algorithm is that their global convergence can be guaranteed and one does not have to check if the estimate given by them is the NPMLE.

As discussed before, for interval-censored failure time data, one can only estimate the probability mass over intervals $(s_{j-1}, s_j]$, but not how the probability is distributed within the interval assuming $\hat{p}_j$ is not zero. In this regard, it has been assumed that $\hat{F}$ puts probability mass only at the s_j's. Sometimes it may be reasonable to assume that F is smooth and thus one may want $\hat{F}$ to be a smooth function rather than a step function over intervals $(s_{j-1}, s_j]$. One way to obtain such an estimate is to apply the self-consistency algorithm to F directly replacing the equation in step 2 of the algorithm by

$$\hat{F}^{(l)}(t) = \frac{1}{n} E\left[\sum_{i=1}^{n} I(T_i \leq t) \,|\, \hat{F}^{(l-1)}, \boldsymbol{O} \right] ,$$

which is based on equation (3.5). In this case, one should choose a strictly increasing distribution function as an initial estimate of F. Li et al. (1997) show that this algorithm converges and gives the same estimated values of F at the s_j's as the self-consistency algorithm described in Section 3.4.1.

Another way to obtain a smooth estimate of a survival or distribution function is to apply smoothing estimation techniques. For example, one can apply the kernel estimation approach to the NPMLE of F or estimate F by obtaining a smooth estimate of the log density function (Pan, 2000c). Similarly, one can obtain smooth estimates of a survival or distribution function using smoothing estimates of the hazard function given in the next section.

3.5 Smooth Estimation of Hazard Functions

Smoothing estimation procedures are commonly used for estimation of hazard functions. This is because nonparametric estimation without smoothing usually gives estimated hazard functions that vary dramatically and are not useful for graphical presentation. For right-censored failure time data, there exist a number of approaches that produce smooth estimates of hazard functions. The simplest approaches are perhaps kernel-based approaches that smooth raw nonparametric estimates using kernel smoothing functions (Lawless, 2003; Tanner and Wong, 1983). Another class of methods uses spline functions to approximate hazard functions (Kooperberg and Stone, 1992; Rosenberg, 1995). In this case, hazard functions are commonly expressed as linear functions of some spline functions. A third type of approaches often used in practice applies local likelihood methods (Tibshirani and Hastie, 1987) that approximate a hazard function at each time point by some parametric functions such as linear functions of time. Penalized likelihood methods are another choice for obtaining smooth estimates.

In the following, we focus attention on the kernel-based approach for smooth estimation of hazard functions based on interval-censored failure time data. This approach has advantages that it is straightforward and can be easily implemented. After two illustrative examples, discussion is provided about some likelihood-based approaches.

3.5.1 Kernel-based Estimation

As in the previous sections, this section assumes that a set of interval-censored data is observed and given by

$$\boldsymbol{O} \,=\, \{\, (L_i, R_i]\,;\, i \,=\, 1, ..., n \,\}\,.$$

Let S, F, $\{\, s_j \,\}_{j=0}^{m}$, α_{ij} and $\boldsymbol{p}$ be defined as before. Also let $\hat{\boldsymbol{p}} \,=\, (\hat{p}_1, ..., \hat{p}_m)'$ denote the maximum likelihood estimator of $\boldsymbol{p}$ and $\lambda(t)$ the hazard function of the underlying failure time.

For estimation of $\lambda(t)$, note that if right-censored data are available, a natural estimate at time s_j is given by d_j/r_j, where d_j and r_j are the observed failure and risk numbers of subjects at s_j, respectively. This suggests that for the current situation, one can estimate $\lambda(t)$ at s_j by

$$\hat{\lambda}_j \,=\, \frac{d_j(\hat{\boldsymbol{p}})\,\hat{p}_j}{\sum_{u=j}^{m}\, d_u(\hat{\boldsymbol{p}})\,\hat{p}_u}$$

given $\hat{\boldsymbol{p}}$, where $d_j(\boldsymbol{p})$ is given in (3.2), $j \,=\, 1, ..., m$. The numerator and denominator in the estimate given above represent, respectively, estimated numbers of failures and risks at s_j. To smooth this raw nonparametric estimate, one can use the moving average approach. More generally, one can apply the kernel estimation method described below (Eubank, 1999).

Let $K(t)$ be a nonnegative function symmetric about $t \,=\, 0$ and suppose that $\int_{-\infty}^{\infty} K(t)\, dt \,=\, 1$. It is usually referred to as a kernel function. Also let h be a positive parameter called the bandwidth parameter, which determines how large a neighborhood of t is used to calculate the local average. Define

$$w_j^*(t, h) \,=\, h^{-1}\, K\{\, (t - s_j)\,/h\,\}$$

and

$$w_j(t) \,=\, \frac{w_j^*(t, h)}{\sum_{u=1}^{m}\, w_u^*(t, h)}\,,$$

$j \,=\, 1, ..., m$. A kernel estimate of $\lambda(t)$ is then given by

$$\hat{\lambda}(t) \,=\, \sum_{j=1}^{m} w_j(t)\,\hat{\lambda}_j\,.$$

In practice, many kernel functions can be used. One simple choice is

$$K_1(t) \,=\, I(\,|t| \,\leq\, 1\,)$$

and under this kernel function, $\hat{\lambda}(t)$ is the moving average estimate. At time t, only these $\hat{\lambda}_j$'s for which $|s_j \,-\, t| \,\leq\, h$ contribute to $\hat{\lambda}(t)$. That is, $\hat{\lambda}(t)$ is the average of the $\hat{\lambda}_j$'s whose corresponding s_j's are within the interval $[t - h, t + h]$. Another kernel function is

$$K_2(t) = (2\pi)^{-1/2} \exp(-t^2/2) ,$$

which is commonly referred to as the Gaussian kernel. Under this function, every $\hat{\lambda}_j$ contributes to $\hat{\lambda}(t)$ and the closer the corresponding s_j to t, the larger the contribution to the $\hat{\lambda}(t)$. More comments about these two kernel functions are given in the next subsection through examples. Once a kernel function K is given, one needs to choose the bandwidth parameter h and for this, one can apply methods commonly used for kernel estimation of density functions (Bean and Tsokos, 1980; Wand and Jones, 1995). Suppose that the goal is to provide a simple, graphical presentation of the hazard function. In this case, the trial and error method seems to be a natural choice. It is obvious that h cannot be too small or large, and the appropriate range for h depends on specific problems.

A major advantage of the kernel estimation approach is its simplicity and flexibility. It can be easily implemented once $\hat{\boldsymbol{p}}$ is obtained and does not need much extra work. A disadvantage of the approach is that it is not likelihood-based and thus inference is not straightforward.

3.5.2 Two Examples

To illustrate the kernel estimation procedure described in the previous subsection, we apply it to two sets of interval-censored data. The first one is the lung tumor data given in Table 1.3 and discussed in Sections 1.2.1 and 3.2, while the other is the breast cancer data presented in Table 1.4 and discussed in Sections 1.2.2 and 3.4.4.

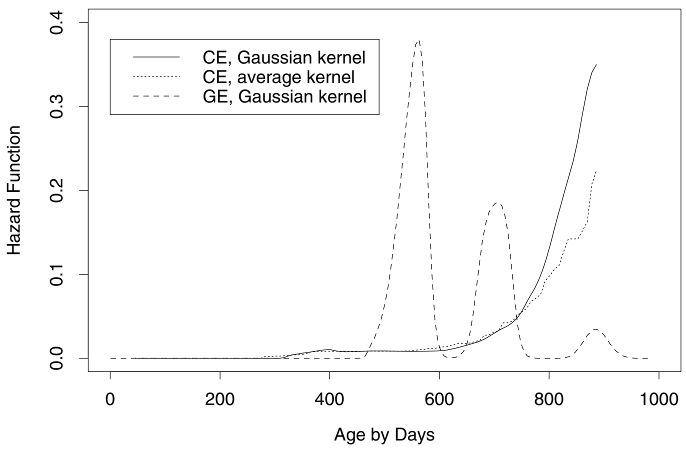

Fig. 3.4. Smooth estimates of hazard functions for time to lung tumor onset.

First we consider estimation of the hazard functions for time to lung tumor onset for animals in each of the two treatment groups, conventional environment (CE) and germ-free environment (GE). Figure 3.4 displays the smooth estimates obtained separately by applying the estimation procedure to each of the two parts of the data. The figure includes three estimates: the estimate with the average kernel K_1 and bandwidth $h = 100$ for animals in the CE group, the estimate with the Gaussian kernel K_2 and bandwidth $h = 20$ also for animals in the CE group, and the estimate with the Gaussian kernel K_2 and bandwidth $h = 20$ for animals in the GE group. It is apparent from the figure that the animals in the two groups have quite different hazard functions, indicating that they have different tumor occurrence rates as suggested by Figure 3.1. In particular, it seems that for the animals in the CE group, the tumor risk increases with time, while for the animals in the GE group, the highest tumor risk occurs roughly between 400 and 800 days.

For the animals in the GE group, only the estimate based on the Gaussian kernel is given. This is because the raw estimates $\hat{\lambda}_j$'s of the hazard function for these animals is quite rough, and, as expected, the average kernel does not seem to produce a good estimate. In other words, for small h, the resulting estimate would not improve much compared with the raw estimate in terms of smoothing, while for large h, the resulting estimate tends to be flat. In contrast, for the animals in the CE group, the raw estimate of the hazard function does not jump up and down very much, and, in fact, its shape is close to the two estimates given in the figure. For these situations, Figure 3.4 suggests that the two kernel functions give similar estimates. The values of

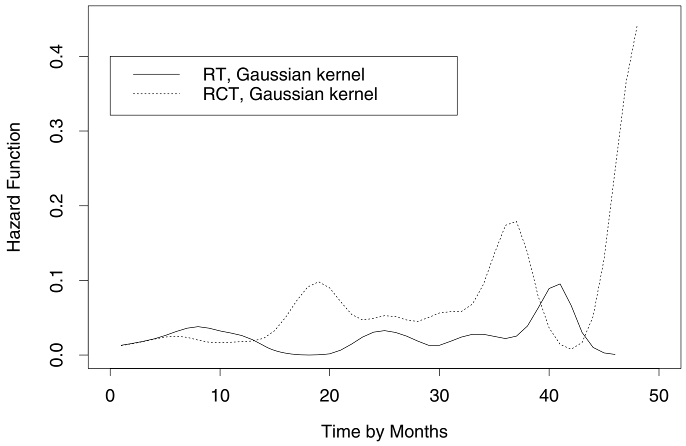

Fig. 3.5. Smooth estimates of hazard functions for time to breast retraction.

the bandwidth h used in the figure are chosen based on the trial and error method.

Figure 3.5 presents the smooth estimates of the hazard functions for time to breast retraction obtained by separately applying the kernel estimation procedure to the two parts of the breast cancer data, the data from patients in the RT group and from those in the RCT group, respectively. For this data set, only the estimates based on the Gaussian kernel are given because as it can be seen from the figure that the raw estimates $\hat{\lambda}_j$'s are quite rough for both groups. For both estimates, the bandwidth used is 2 and is selected using the same way as with Figure 3.4. As Figure 3.2, Figure 3.5 seems to indicate that the patients in the two treatment groups have different breast retraction rates with the patients in the RCT group having higher risk to develop breast retraction than those in the RT group. Also for the patients in the RT group, the risk does not seem to change much, while for those in the RCT group, the risk can be lower for some time periods and higher for some other time periods.

3.5.3 Likelihood-based Approaches

In addition to the kernel-based approaches, for smooth estimation of hazard functions, one can use some likelihood-based approaches such as penalized likelihood methods and local likelihood methods. Let $\Lambda(t) = \int_0^t \lambda(s)\, ds$, the cumulative hazard function. Define $\delta_i = I(L_i = R_i)$, indicating (by 1) if the observation on subject i is exact. Then the likelihood function in (3.1) has the form

$$L_S(\,\lambda(t)\,) = \prod_{i=1}^{n} \left[\lambda(L_i)\, e^{-\Lambda(L_i)}\right]^{\delta_i} \left[e^{-\Lambda(L_i)} - e^{-\Lambda(R_i)}\right]^{1-\delta_i}$$

and one may estimate $\lambda(t)$ by maximizing $L_S(\,\lambda(t)\,)$. However, as with estimation of survival functions, the maximization of this likelihood function only gives estimates of the hazard function at discrete time points. More seriously, the resulting estimate is usually not smooth, while it often may be reasonable to assume that $\lambda(t)$ is a smooth function.

To obtain a smooth estimate of $\lambda(t)$, one can use the penalized likelihood approach, which maximizes the log likelihood function adjusted by a penalty function such as

$$l_p(\,\lambda(t)\,|\,\tau) = l_S\,(\,\lambda(t)\,) - \frac{\tau}{2}\, g\,(\,\lambda(t)\,)\,. \tag{3.8}$$

In $l_p(\,\lambda(t)\,|\,\tau)$, $l_S(\,\lambda(t)\,) = \log L_S(\,\lambda(t)\,)$, g is a known penalty function measuring the roughness of the hazard function, and τ (>0) is an unknown parameter that controls the amount of smoothing. If $\tau = 0$, $l_p(\,\lambda(t)\,|\,\tau) = l_S(\,\lambda(t)\,)$ and there is no smoothing. The penalized likelihood approach aims to balance smoothness of the hazard function against its fit to the observed data.

Suppose that $\lambda(t)$ is a smooth function. In this case, a common approach to impose this smoothness is to express the log hazard function as

$$\log \lambda(t) \,=\, \sum_{j=0}^{p} \alpha_j \, B_j(t) \;, \tag{3.9}$$

where $\{\, \alpha_j \,;\, j = 1, ...p \,\}$ are unknown parameters and $\{\, B_j(t) \,;\, j = 1, ...p \,\}$ are some known smooth functions. For estimation of $\lambda(t)$, instead of maximizing the penalized log likelihood function given in (3.8), one can maximize it with $\lambda(t)$ replaced by model (3.9) and in this case, $l_S(\, \lambda(t)\,)$ can be written as

$$l_S(\alpha_j' s) \,=\, \sum_{i=1}^{n} \left\{ \delta_i \left[\sum_{j=0}^{p} \alpha_j \, B_j(L_i) - \int_0^{L_i} \exp\left(\sum_{j=0}^{p} \alpha_j B_j(s) \right) ds \right] \,+\, (1 - \delta_i) \right.$$

$$\left. \times \,\log \left[\exp\left(- \int_0^{L_i} e^{\sum_{j=0}^{p} \alpha_j B_j(s)} ds \right) \,-\, \exp\left(- \int_0^{R_i} e^{\sum_{j=0}^{p} \alpha_j B_j(s)} ds \right) \right] \right\} \,.$$

The model (3.8) says that the hazard function $\lambda(t)$ is a linear function of some known functions. There are many choices for functions $B_j(t)$'s. For example, one simple choice is given by taking them to be power functions, in which the log hazard function is a polynomial function of time t. Another choice is to let each $B_j(t)$ be some known spline functions such as B-splines or M-splines, or base functions of some function space. Cai and Betensky (2003) proposed to use

$$\log \lambda(t) \,=\, \alpha_0 \,+\, \alpha_1 \, t \,+\, \sum_{j=1}^{p} \beta_j \, (t \,-\, t_j)^+ \,,$$

where $a^+ \;=\; \max(0, a)$ and $\{\, t_j \,;\, j \,=\, 1, ..., p \,\}$ are preselected time points commonly referred to as knots in the spline literature. That is, $\lambda(t)$ follows a linear spline model. In this case, by using the quadratic penalty function, the penalized log likelihood function has the form

$$l_p(\boldsymbol{\theta} \,|\, \sigma^2) \,=\, l_S(\boldsymbol{\theta}) \,-\, \frac{\tau}{2} \boldsymbol{\beta}' \, \boldsymbol{\beta} \,,$$

where $\boldsymbol{\theta} \,=\, (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, $\boldsymbol{\alpha} \,=\, (\alpha_0, \alpha_1)'$, and $\boldsymbol{\beta} \,=\, (\beta_1, ..., \beta_p)'$.

Instead of modeling the log hazard function, one can also directly model the hazard function. For example, Joly et al. (1998) suggest using M-splines to model the hazard function. Rosenberg (1995) discusses the same idea under the model

$$\lambda(t) \,=\, \sum_{j} B_j(t) \, \exp(\alpha_j) \,.$$

The methods discussed above are in fact a combination of the penalized likelihood approach and the modeling of the hazard function. For smooth estimation of $\lambda(t)$, instead of using this combination method, one could obtain the estimates by maximizing the penalized log likelihood function $l_p(\lambda(t) \mid \tau)$ without using model (3.9). Bacchetti (1990) applies this approach to the hazard function for AIDS incubation time. Similarly, one can directly maximize the log likelihood function $L_S(\, \alpha_j' s \,)$ under model (3.9) (Kooperberg and Clarkson, 1997). This corresponds with the weakly parametric estimation techniques or models discussed in Section 2.5. It should be noted, however, that the weakly parametric estimation approach itself may yield estimates that are not very smooth when p is large.

The local likelihood method was proposed by Tibshirani and Hastie (1987) for smooth estimation of covariate effects in the context of regression analysis. It is likelihood-based and an extension of the local fitting technique used in scatterplot smoothing (Cleveland, 1979). Betensky et al. (1999) apply the method to estimation of the hazard function. To implement the method, a set of intervals has to be preselected and in each interval, the hazard function is approximated by a linear function of time. The parameters in the linear function are estimated using the local likelihood contributed by the interval-censored failure times related to the interval over which the linear model is defined. The approach also needs to choose the parameter controlling the amount of smoothing. A drawback of this approach is that it could have numerical stability problems in the regions of sparse data.

3.6 Asymptotics

This section deals with asymptotic properties of the NPMLE of S or F. As seen in Sections 3.2 to 3.4, for interval-censored data, determination of the NPMLE is much more complex and difficult than that for right-censored data. This is mainly because the amount of information about the failure time variable of interest given by the former is much less than that given by the latter. The same phenomenon also makes asymptotic properties of the NPMLE much harder to study and quite different from those of the NPMLE given by right-censored data. In the case of right-censored data, the NPMLE can be conveniently expressed using counting processes (Andersen et al., 1993). Hence martingale theory is readily available for the study of asymptotic properties of the NPMLE such as consistency and asymptotic normalityindexAsymptotic properties!asymptotic distribution. In contrast, for interval-censored data, the same is not true anymore due to the structure of the data.

Let T denote the failure time variable of interest and $\hat{F}_n(t)$ the NPMLE of the distribution function of T based on a set of interval-censored failure time data from n independent individuals. Also, let F_0 denote the underlying true distribution function of T. To discuss the asymptotic properties of $\hat{F}_n(t)$, one needs to separate current status data and case II interval-censored because they have different structures and thus require different types of conditions. For current status data, as before, let C denote the observation time variable that is assumed to be independent of T and G its distribution function. Then for each subject under study, only C and the indicator function $I(T \leq C)$ are observed.

For case II interval-censored data, we need to use the formulation (1.2) for asymptotic studies. In this case, let H and h denote the joint distribution and density functions of observation time variables U and V, respectively, assumed to be independent of T. Also let H_1, H_2 and h_1, h_2 denote the marginal distribution and density functions of U and V, respectively. In the following, three basic asymptotic properties of $\hat{F}_n(t)$ are discussed. They are its consistency, local asymptotic distribution of $\hat{F}_n(t)$, and the asymptotic distribution of linear functionals of $\hat{F}_n(t)$. For the asymptotic distributions of both $\hat{F}_n(t)$ and its linear functionals, a distinction needs to be made in terms of the length of observed intervals or the mass of H along the diagonal.

3.6.1 Consistency

Consistency is perhaps the most fundamental property that one investigates for an estimate. Assume that F_0 is continuous. Then for interval-censored failure time data, Groeneboom and Wellner (1992) show that $\hat{F}_n(t)$ is uniformly consistent in the sense that

$$P\left\{ \lim_{n \to \infty} \sup_{t \in \mathcal{R}^+} \left| \hat{F}_n(t) - F_0(t) \right| \right\} = 1 . \tag{3.10}$$

This uniform consistency is true for both current status data and case II interval-censored data.

Some regularity conditions are needed for the uniform consistency given above. For current status data, the proof of (3.10) assumes that like F_0, G is also continuous and requires, among others, the following condition.
(C1). $F_0 \ll G$. That is, if G puts zero mass on a set A, then F_0 has zero mass on A as well.

For case II interval-censored data, a condition similar to (C1) is required. Specifically,
(I1). $F_0 \ll H_1 + H_2$. That is, if both H_1 and H_2 put zero mass on a set A, then F_0 has zero mass on A too.

Both conditions (C1) and (I1) mean that F_0 has probability zero on sets in which no observation can occur. In other words, for the NPMLE of F_0 to be consistent, the support of F_0 needs to be contained in the support of G or in the union of the supports of H_1 and H_2.

Assume that the support of all F_0, G, H_1 and H_2 is a bounded interval $I = [0, M]$ with $M > 0$. Then one can also establish the L_2-consistency of $\hat{F}_n(t)$ given by

$$||\hat{F}_n - F_0||_G = \int_0^M \left[\hat{F}_n(t) - F_0(t) \right]^2 dG(t) = O_p(n^{-1/3}) \tag{3.11}$$

for current status data (Huang and Wellner, 1995), or

$$||\hat{F}_n - F_0||_{H_i} = \int_0^M \left[\hat{F}_n(t) - F_0(t)\right]^2 \, dH_i(t) = O_p(n^{-1/3}) \qquad (3.12)$$

for $i = 1, 2$ for case II interval-censored data (Geskus and Groeneboom, 1997).

In addition to the consistency, another fact implied by (3.11) or (3.12) is that the NPMLE of F_0 based on interval-censored data only has $n^{1/3}$-convergence rate in L_2 measure. This is quite different from $n^{1/2}$-convergence rate of the NPMLE based on right-censored data. Other authors who investigated the consistency problem include Schick and Yu (2000) and Yu, Li, and Wong (2000). The former established the L_1-consistency of the NPMLE and the latter considered the strong consistency of the self-consistent estimate of F when the support of the observation times is finite.

3.6.2 Local Asymptotic Distribution

The local asymptotic distribution or asymptotic distribution of $\hat{F}_n(t)$ at a given time point t_0 differs for current status data and case II interval-censored data. It also depends on if the observed data contain exact or nearly exact failure times. As in the previous subsection, assume that F_0 is continuous.

For current status data, assume that $0 < F_0(t_0)\,,\, G(t_0) < 1$ and that F_0 and G have density functions f_0 and g, respectively, that are strictly positive at t_0. Then Groeneboom and Wellner (1992) show that as $n \to \infty$,

$$n^{1/3} \left[\frac{2\, g(t_0)}{f_0(t_0)\, F_0(t_0)\, [1 - F_0(t_0)]}\right]^{1/3} \left[\hat{F}_n(t_0) - F_0(t_0)\right] \to 2\, Z \qquad (3.13)$$

in distribution, where Z is the last time where standard Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

Two aspects about $\hat{F}_n(t)$ can be immediately seen from (3.13). One is that as mentioned before, $\hat{F}_n(t)$ has $n^{1/3}$-convergence rate rather than the $n^{1/2}$-convergence rate that usually holds for estimates based on right-censored data The other is that $\hat{F}_n(t)$ does not have an asymptotic normal distribution as one may expect for an NPMLE.

For case II interval-censored data, one can derive an asymptotic result that is similar to (3.13), but needs stronger conditions. In particular, assume that both F_0 and G have support $[0, M]$, a bounded interval, and f_0 satisfies

$$f_0(t) \geq a_0 > 0\,,\; t \in (0, M)\,,$$

for some constant $a_0 > 0$. Also one needs:
(I2). h_1 and h_2 are continuous with $h_1(t) + h_2(t) > 0$ for all $t \in [0, M]$.
(I3). $P(\, V - U < \epsilon\,) = 0$ for some positive ϵ.

Condition (I2) says that any t within $(0, M)$ can be an observation time point with positive likelihood, while condition (I3) implies that the joint density function h does not have mass close to the diagonal or the joint distribution function H has zero mass on some strip along the diagonal. In terms of the observed data, the condition (I3) means that all observed intervals are real intervals and no exact failure times are available. Under the conditions given above plus some other regularity conditions, one has that as $n \to \infty$,

$$n^{1/3} \left[\frac{2a(t_0)}{f_0(t_0)}\right]^{1/3} \left[\hat{F}_n(t_0) - F_0(t_0)\right] \to 2\,Z \tag{3.14}$$

in distribution for each point $t_0 \in (0, M)$ (Groeneboom, 1996). In (3.14), Z is defined as in (3.13) and

$$a(t_0) = \frac{h_1(t_0)}{F_0(t_0)} + \frac{h_2(t_0)}{1 - F_0(t_0)} + \int_{t_0}^{M} \frac{h(t_0, v)}{F_0(v) - F_0(t_0)} dv + \int_0^{t_0} \frac{h(u, t_0)}{F_0(t_0) - F_0(u)} du \,.$$

The asymptotic distribution given in (3.14) says that as with current status data, the NPMLE from case II interval-censored data also has $n^{1/3}$-convergence rate and does not have an asymptotic normal distribution. For both (3.13) and (3.14), it is assumed that no exact failure times are observed or available. Suppose that some exact failure times are available and let n_1 denote the number of subjects for whom such exact failure times are observed. Define condition (CI1) as

(CI1). $n_1/n \to b_0$ as $n \to \infty$ with b_0 a positive constant.

Huang (1999b) shows that under condition (CI1) and some other regularity conditions as $n \to \infty$, one has

$$n^{1/2} \left[\hat{F}_n(t_0) - F_0(t_0)\right] \to Z_1(t_0) \tag{3.15}$$

in distribution. Here $Z_1(t_0)$ is a normal random variable with mean zero and the variance given by the information lower bound for the estimation of F_0. The result given above is true for both current status data and case II interval-censored data, but the conditions required for the former are weaker than those for the latter. For instance, in the case of current status data, no restriction is needed for G, whereas some restrictions similar to (I3) are required for observation times for case II interval-censored data.

The result given in (3.15) says that the NPMLE of F_0 can have $n^{1/2}$-convergence rate and an asymptotic normal distribution if enough exact failure times or more information about F_0 than purely interval-censored observations is available. For situations that lie between conditions (I3) and (CI1), one can expect a convergence rate that is between $n^{1/3}$ and $n^{1/2}$. For example, Groeneboom and Wellner (1992) discuss a case where H has sufficient mass along the diagonal and suggest that $\hat{F}_n$ has $(n \log n)^{1/3}$-convergence rate.

One application of (3.13) - (3.15) is to construct pointwise confidence limits for F_0. For this and the situations corresponding to (3.13) and (3.14), one has to estimate the constants at the left side of (3.13) and (3.14) in addition to others. The estimation of these constants involves estimation of density functions, which may not be an easy task. For the situation corresponding to (3.15), one needs to estimate the variance of $\hat{F}_n$ because its asymptotic variance does not have a closed form. A consistent estimate of the asymptotic variance is given by the observed information matrix, the negative second derivative of the log-likelihood. In this case, based on (3.15), one can also apply the bootstrap procedure to construct confidence bands for F_0 (Huang, 1999b). In general, it should be easier to use likelihood ratio statistics to construct confidence limits of F_0 at a fixed time point if their asymptotic distributions are known.

In the preceding discussion, it is assumed that F_0 is a continuous distribution function. Sometimes it is of interest and reasonable to treat F_0 as a discrete distribution function with finite known support points (Yu et al., 1998a, b; Yu, Wong and Li, 2001). Then the derivation of the asymptotic consistency and distribution of the NPMLE becomes a standard parametric problem.

3.6.3 Asymptotic Normality of Linear Functionals

This subsection discusses the asymptotic distribution of linear functionals given by

$$K(\hat{F}_n) = \int c(t)\, d\hat{F}_n(t)\,,$$

where $c(t)$ is a given function. Taking $c(t) = t$, one has $K(F_0) = E(T)$, the mean of failure time variable of interest. As for the local asymptotic distribution of $\hat{F}_n$, one has to separate the investigation of the asymptotic distribution of $K(\hat{F}_n)$ for current status data and case II interval-censored data. In the following, it is assumed that F_0 has bounded support, i.e., $[0, M]$ with $M > 0$.

First for current status data, assume that $G \ll F_0$, $F_0 \ll G$ and G has a density function g. One can show that

$$\sqrt{n}\,\left[K(\hat{F}_n) - K(F_0)\right] \to N(0, \sigma_c^2) \tag{3.16}$$

in distribution as $n \to \infty$, where

$$\sigma_c^2 = \int_0^M \frac{F_0(t)\,[1 - F_0(t)]}{g(t)}\,[c'(t)]^2\,dt\,,$$

which is assumed to exist (Groeneboom and Wellner, 1992; Huang and Wellner, 1995). This result tells us that although $\hat{F}_n$ only has $n^{1/3}$-convergence rate, its linear functionals have the usual $n^{1/2}$-convergence rate.

For case II interval-censored data, one can prove that (3.16) still holds, but there is no explicit formula available for the asymptotic variance σ_c^2. In fact, Geskus and Groeneboom (1997, 1999) show that (3.16) is true for general smooth functionals of $\hat{F}_n$ and that the asymptotic variance reaches the information lower bound. As for the local asymptotic distribution of $\hat{F}_n$ discussed in the previous subsection, to have (3.16), one does need more and stronger conditions about the observation process. In particular, one has to pay special attention to the behavior of H along the diagonal. Geskus and Groeneboom (1997) considered situations where conditions (I1) and (I2) are true. That is, H has zero mass along the diagonal. Geskus and Groeneboom (1999) dealt with situations where U and V can be arbitrarily close, i.e., the following condition holds:

(I4). $h(t,t) = \lim_{v \downarrow t} h(t,v) \geq c_0 > 0$ for all $t \in (0, M)$ and some $c_0 > 0$.

3.7 Bibliography, Discussion, and Remarks

Nonparametric maximum likelihood estimation is the topic that has been discussed the most in the analysis of interval-censored failure time data, and the study of it goes back to the fifties. Ayer et al. (1955) and Eeden (1956) were the first to derive the NPMLE of a distribution function based on current status data. Peto (1973) and Turnbull (1976) investigated the estimation problem based on general or case II interval-censored data. For the problem, the former presented a Newton-Raphson algorithm and the latter developed the self-consistency algorithm described in Section 3.4.

Following the seminal article Turnbull (1976), a number of authors have studied the nonparametric estimation problem for interval-censored data from various points of view. For example, as pointed out in the previous sections, Gentleman and Geyer, 1994, Groeneboom and Wellner (1992), Jongbloed (1998), Li et al. (1997) and Wellner and Zhan (1997) discussed issues related to determination of the NPMLE based on interval-censored data such as its characterization and algorithms. Others that dealt with similar or related issues include Banerjee and Wellner (2005), Becker and Melbye (1991), Böhning et al. (1996), Braun et al. (2005), Frydman (1994), Goodall et al. (2004), Groeneboom (1995), Hudgens (2005), Hudgens and Satten (2001), Ng (2002), Pan (2000c), Pan and Chappell (1998a), Rücker and Messerer (1988), Song (2004), Sun (2001a), Vandal et al. (2005), van der Laan and Jewell (2003), and Yu et al. (1998a). Among them, Banerjee and Wellner (2005) and Goodall et al. (2004) discussed the construction of confidence limits or intervals for current status data and case II interval-censored data, respectively. Böhning et al. (1996) noted the similarity between the problem considered here and the finite mixture model estimation problem and suggested using the vertex-exchange or other algorithms proposed for the latter situations to determine the NPMLE. Braun et al. (2005) and Pan (2000c) considered smooth estimation of a density or survival function, respectively, and Frydman (1994), Hudgens (2005) and Pan and Chappell (1998a) studied the estimation problem when truncation is present as well as interval censoring. Hudgens and Satten (2001) discussed interval censoring involved in competing risk data, Sun (2001a) considered pointwise variance estimation of the NPMLE, and Vandal et al. (2005) studied the constrained nonparametric estimation problem.

As discussed in Section 3.6, the study of asymptotic properties of the NPMLE based on interval-censored failure time data is a difficult, but important issue. For this, much of the important ground work is laid out in Groeneboom and Wellner (1992) and Groeneboom (1996) and one needs to use empirical process theory (van der Vaart and Wellner, 1996). For the consistency of the NPMLE, the authors who studied it include Ayer et al. (1955), Geskus and Groeneboom (1997), Groeneboom and Wellner (1992), Huang and Wellner (1995, 1997), Pan and Chappell (1998b), Pan et al. (1998), Schick and Yu (2000), van de Geer (1993), Wang and Gardiner (1996), Yu, Li, and Wong (2000), Yu et al. (1998a, b), and Yu et al. (2000). In particular, Pan and Chappell (1998b) and Pan et al. (1998) dealt with data that involve both interval censoring and left-truncation. The references that discussed the asymptotic distribution of the NPMLE or its functionals include Geskus and Groeneboom (1996, 1997, 1999), Groeneboom (1996), Groeneboom and Wellner (1992), Huang (1999b), Huang and Wellner (1995, 1997), and Yu et al. (1998a).

Several problems that are closely related to nonparametric estimation of a distribution function are not discussed in the preceding sections. One is constrained nonparametric estimation of a distribution function based on interval-censored failure time data such as the maximization of the log likelihood functions discussed in the previous sections under constraints. There exist many situations where one needs to use the procedures developed for such constrained maximization. For instance, one may want to construct likelihood intervals for the NPMLE of a distribution function. Sometimes, it may be reasonable to assume that the true distribution function F_0 is concave, unimodal, has a monotone hazard function, etc. For the problem, Vandal et al. (2005) gave a reduction technique as well as some algorithms for determination of the NPMLE under constraints for interval-censored data.

In addition to interval censoring, truncation may exist, and one may desire a nonparametric estimate of a distribution function that takes into account both interval censoring and truncation. The original self-consistency algorithm given in Turnbull (1976) covers this situation, and as mentioned before, Pan and Chappell (1998a, b) also discussed the situation.

Sometimes one may be interested in smooth estimation of a density function based on interval-censored failure time data. A simple approach uses the NPMLE of a distribution function discussed in the previous sections. For example, a simple kernel estimate is determined by

$$\int \frac{1}{n\,h} \sum_{i=1}^{n} E_{\hat{F}_n} \left[K \left(\frac{t - T_i}{h} \right) | \, T_i \in (L_i, R_i] \right]$$

given a set of interval-censored data $\{(L_i, R_i]\,;\, i = 1, ..., n\}$, the NPMLE $\hat{F}_n$, a kernel function K and a bandwidth h. In the formula above, $E_{\hat{F}_n}$ denotes the conditional expectation of T_i, the unobserved failure time from subject i, with respect to $\hat{F}_n$ given $\hat{F}_n$ and the observed interval $(L_i, R_i]$ for T_i. One can also obtain a smooth estimate of a density function using the smooth estimate of a hazard function given in Section 3.5 and the relationship between hazard and density functions discussed in Section 1.4.1. Braun et al. (2005) proposed a third approach, which incorporates kernel estimation into self-consistency and local likelihood procedures.

Also estimation of the variance or covariance of the NPMLE of a distribution function was not considered. Such estimates are needed in order, for example, to construct confidence bands of a distribution function. If it is reasonable to treat the distribution function as a discrete function with finite support, then one can easily obtain an estimate of the covariance matrix of the NPMLE using the observed information matrix. In case the number of time points in the support is large, this method could be computationally complex and give unrealistic results. For this, other available methods for interval-censored data in the literature include the likelihood ratio approach considered in Banerjee and Wellner (2005) and Goodall et al. (2004) for confidence intervals and the two sampling-based approaches given in Sun (2001a) for pointwise variance estimation of the NPMLE. For the two sampling-based approaches, one is a simple bootstrap procedure and the other is a generalized Greenwood's formula based on resampling. Also one can use the consistent estimate of the asymptotic variance given in Yu, Li, and Wong (1998).

Although studies have been conducted that deal with asymptotic properties of the NPMLE based on interval-censored failure time data, there still exist many open questions about asymptotic behaviors of the NPMLE. For instance, the local asymptotic distribution of the NPMLE under condition (I4) is still unknown. Another unresolved, interesting question concerns the local asymptotic distribution of a constrained NPMLE. By imposing some constraints on the distribution function, one may be able to estimate it with a better convergence rate. Also, it would be useful to investigate the asymptotic distributions of likelihood ratio statistics, which can be used, for example, to construct confidence limits of F_0.

4

Comparison of Survival Functions

4.1 Introduction

Comparison of treatments is one of the primary objectives in most medical studies such as clinical trials. In such cases, nonparametric or distribution-free methods are usually preferred if there does not exist strong evidence to support a particular parametric model. For right-censored failure time data, most of the existing nonparametric methods can be classified into two types: weighted log-rank tests and weighted Kaplan-Meier . In particular, the log-rank test is perhaps the most commonly used nonparametric procedure in practice. Detailed discussions about these two types of statistics can be found in Fleming and Harrington (1991) and Kalbfleisch and Prentice (2002) among other books. This chapter deals with similar methods that are appropriate for interval-censored failure time data. Alternatives to these methods, which are discussed in Chapters 5 and 6, base the comparisons on the score tests derived under various regression models.

In Section 4.2, we discuss nonparametric treatment comparisons when only current status data, i.e., case I interval-censored data, are available and two types of methods are considered. One approach is appropriate only if observations times across treatment groups follow the same distribution and the other allows different distributions for these observation times. Sections 4.3 and 4.4 consider, respectively, rank-based and survival-based comparison methods for general or case II interval-censored data. In both sections, we concentrate on situations where the censoring intervals for subjects in different treatment groups are generated from the same distribution function. The use of these methods are illustrated in Section 4.5 on three examples including the lung tumor data described in Section 1.2.1 and the breast cancer data described in Section 1.2.2. Section 4.6 includes bibliographic notes about nonparametric treatment comparison for interval-censored data and brief discussion about situations and approaches that are not considered in the previous sections.

4.2 Statistical Methods for Current Status Data

As before, we use T to denote a random variable representing failure time and S its survival function. Let C denote the observation time that is assumed to be independent of T. Suppose that data are observed independently from n subjects, and for each subject, one observes only C and $\delta = I(T \leq C)$. For treatment comparisons, we consider two separate situations because the methods for them are quite different. One assumes that C follows the same distribution for all subjects, and the other allows the distributions of C to be different for subjects with different treatments.

4.2.1 Comparisons with the Same Observation Time Distribution

In this subsection, we focus on the two-sample comparison problem and discuss three simple and natural methods. The first one is a Wilcoxon-type procedure, and the second one is a rank-based method first developed in Sun and Kalbfleisch (1993, 1996) following the idea behind the log-rank test. The third approach is a generalization of the weighted Kaplan-Meier procedure, and the idea behind it can be found in Andersen and Ronn (1995) among others. With S_1 and S_2 denoting the survival functions corresponding with the two treatment groups, the goal is to test the hypothesis $H_0 : S_1(t) = S_2(t)$ for all t or $S_1 = S_2$.

4.2.1.1 A Wilcoxon-type Procedure

Suppose that the observed data consist of $\{(C_i\,,\,\delta_i\,,\,Z_i)\,;\,i = 1, ..., n\}$, where $Z_i = 0$ or 1 is the treatment group indicator for subject i. To test H_0, we note that under H_0 and the assumption that the C_i's follow the same distribution, the δ_i's are i.i.d. random variables. This suggests that one can use the following Wilcoxon statistic

$$\sum_i \sum_j (Z_i - Z_j)\,(\delta_i - \delta_j)$$

for testing H_0. It is apparent that the above statistic is equivalent to

$$U_{cw} = \sum_{i=1}^{n} (Z_i - \bar{Z})\,\delta_i \,,$$

where $\bar{Z} = n^{-1} \sum_{i=1}^{n} Z_i$. It is easy to show that under H_0 and for large n, the distribution of $n^{-1/2}\,U_{cw}$ can be approximated by a normal distribution with mean zero and variance $\hat{\sigma}_{cw}^2 = n^{-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2\,\delta_i^2$. Hence a large sample test of the hypothesis H_0 can be performed using the statistic $U_{cw}/\hat{\sigma}_{cw}$ with standard normal critical values.

4.2.1.2 A Rank-based Procedure

Let $s_1 < ... < s_k$ denote the distinct ordered observation times of $\{ C_1, ..., C_n \}$ and $\hat{S}_0$ the maximum likelihood estimator of the common survival function S_0 under H_0. Also let $F_0(t) = 1 - S_0(t)$ and $\hat{F}_0 = 1 - \hat{S}_0(t)$. Define $n_j = \sum_{\{i\,:\,C_i = s_j\}} \delta_i$, the number of subjects observed to have experienced the survival event at time s_j, and $l_j = \sum_{\{i\,:\,C_i = s_j\}} 1$, the number of subjects observed at time s_j, $j = 1, ..., k$. Then as discussed in Section 3.2, $\hat{F}_0(s_j)$ is given by

$$\hat{F}_0(s_j) = max_{r \leq j}\, min_{s \geq j}\, \frac{\sum_{v=r}^{s} n_v}{\sum_{v=r}^{s} l_v} \quad . \tag{4.1}$$

To test H_0, we note that the log-rank test statistic for right-censored data has the following form: the summation of the observed minus expected failure numbers. This motivates a simple and natural test statistic

$$U_{cr} = \sum_{i=1}^{n} Z_i \, [\, \delta_i - \hat{F}_0(C_i) \,] \,,$$

the summation of the observed minus expected numbers of events over all subjects. This statistic, U_{cr}, can also be derived as a score test statistic from the logistic model

$$\log \frac{1 - S_2(t)}{S_2(t)} = \log \frac{1 - S_1(t)}{S_1(t)} + \beta$$

for testing $\beta = 0$, where β represents the treatment difference. Under H_0, it can be shown that if n is large, the distribution of U_{cr} can be approximated by the normal distribution with mean zero and variance

$$\hat{\sigma}_{cr}^2 = \hat{\sigma}_z^2 \sum_{i=1}^{n} [\, \delta_i - \hat{F}_0(C_i) \,]^2 \,,$$

where $\hat{\sigma}_z^2$ denotes the sample variance of the Z_i's.

The proof for the normal approximation to the distribution of U_{cr} and further discussion are found in Sun and Kalbfleisch (1993, 1996). In fact, the proof requires that the Z_i's can be regarded as i.i.d. variables. However, the finite sample studies presented there suggest that this assumption can be relaxed to the condition that the Z_i's are a random permutation of 0 and 1. This is usually the case for randomized studies.

4.2.1.3 A Survival-based Procedure

Let G denote the distribution function of C and $\hat{G}_n$ the empirical distribution of C. Suppose that both S_0 and G are absolutely continuous functions. Let $\hat{S}_1$ and $\hat{S}_2$ be the maximum likelihood estimators of S_1 and S_2, respectively, as given by (4.1), but based on the subjects within each treatment group separately. Also, let n_1 and n_2 denote the numbers of subjects receiving each of the two treatments, respectively, and assume that $n_1/n \to p$ $(0 < p < 1)$ as $n \to \infty$, where $n_1 + n_2 = n$. To test H_0, motivated by the Kolmogorov-Smirnov and the weighted Kaplan-Meier tests (Pepe and Fleming, 1989), we can construct a simple test statistic

$$U_{cs} = \int_0^{\tau} [\, \hat{S}_2(t) - \hat{S}_1(t)\,]\, d\,\hat{G}_n(t)\,, \tag{4.2}$$

where τ is a constant such that $S_0(\tau) > 0$ and is usually taken to be the longest observation time.

The asymptotic distribution of U_{cs} follows easily from the strong consistency of $\hat{G}_n$ and the results discussed in Section 3.6. Specifically, if n is large and H_0 is true, the distribution of U_{cs} multiplied by $\sqrt{n}$ can be approximated by the normal distribution with mean zero and variance

$$\hat{\sigma}_{cs}^2 = \frac{n^2}{n_1\, n_2} \int_0^{\tau} \hat{S}_0(t)\, [1 - \hat{S}_0(t)]\, d\,\hat{G}_n(t)\,.$$

Thus a large sample test of the hypothesis H_0 can be performed using $U_{cs}^* = \sqrt{n}\, U_{cs}/\hat{\sigma}_{cs}$ with standard normal critical values.

Following the idea behind U_{cs}, similar test statistics can be developed. For example, one can use the statistic U_{cs} with $\hat{S}_1$ and $\hat{S}_2$ replaced by their squares. Another alternative is to consider the difference between the empirical means of the two populations. The methods based on these two statistics, which are expected to perform like that based on U_{cs}, are discussed in detail in Andersen and Ronn (1995) and Tang et al. (1995), respectively.

Simplicity is a key advantage of the methods based on U_{cw}, U_{cr} or U_{cs} which can be implemented easily as t-tests. We remark that as in the case of complete or right-censored failure time data, the first method applies to more general situations, but may have less power for some specific alternatives. For the other two approaches, it is expected that the second method based on U_{cr} should have better power than the third approach based on U_{cs} for alternatives with ordered hazard functions. The survival-based approach would perform better than the rank-based one for alternatives with ordered survival functions. This is discussed further below for general interval-censored data.

4.2.2 Comparisons with Different Observation Time Distributions

This subsection considers nonparametric comparisons of $p + 1$ treatments when observation times may follow different distributions for subjects with different treatments. We mainly focus on the test procedure given in Sun (1999). Let $\boldsymbol{Z}$ be the p-dimensional vector of treatment indicators, and for convenience in presenting the method described below, define $N(t) = I(T \leq t)$, indicating if the survival event has occurred by time t. Then the observed data consist of $\{\, (C_i, N_i(C_i), \boldsymbol{Z}_i) \,;\, i \,=\, 1, ..., n \,\}$, where N_i, C_i, and $\boldsymbol{Z}_i$ are defined as above for subject i, $i \,=\, 1, ..., n$. The hypothesis that the $p \,+\, 1$ treatments are equal is equivalent to the hypothesis $H_0 \,:\, E\{\, N_i(t) \,|\, \boldsymbol{Z}_i\}$ is independent of $\boldsymbol{Z}_i$.

To motivate the approach below, consider the statistic U_{cw} discussed in Section 4.2.1. Using the notation of this subsection, it can be written as

$$\sum_{i=1}^{n} (\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}})\, N_i(C_i)\,. \tag{4.3}$$

Because the distribution of the C_i's may depend on the $\boldsymbol{Z}_i$'s, it is apparent that the statistic in (4.3), which measures the observed treatment differences, may be biased. To correct this possible bias, we need to specify the dependence of the distribution of the C_i's on the $\boldsymbol{Z}_i$'s. For this purpose, we assume that the hazard function of the C_i's is given by the PH model

$$\lambda(t\,;\, \boldsymbol{Z}_i) \,=\, \lambda_0(t)\, e^{\boldsymbol{Z}_i' \boldsymbol{\beta}}\,, \tag{4.4}$$

where $\lambda_0(t)$ denotes an unknown baseline hazard function and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters.

Under this PH model and H_0, it can be shown that

$$E\,[N_i(C_i)|\boldsymbol{Z}_i] \,=\, E\left[\int_0^\infty N_i(t) d\tilde{N}_i(t) | \boldsymbol{Z}_i\right]$$

$$= e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} \int_0^\infty \lambda_0(t) \mu(t) [S_0(t)]^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} dt,$$

where $\mu(t)$ is the mean function of the $N_i(t)$'s under H_0, $\tilde{N}_i(t) \,=\, I(t \geq C_i)$ and $S_0(t) \,=\, \exp[-\int_0^t \lambda_0(s)\, ds]$ is the baseline survival function of the C_i's. This yields

$$E\left[e^{-\boldsymbol{Z}_i' \boldsymbol{\beta}} \int_0^\infty \frac{N_i(t)\, d\tilde{N}_i(t)}{S_0(t)^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})}} \,|\, \boldsymbol{Z}_i\right] \,=\, \int_0^\infty \lambda_0(t)\, \mu(t)\, dt$$

and, assuming that $\boldsymbol{\beta}$ is known, suggests the test statistic

$$\boldsymbol{U}_{cc}(\boldsymbol{\beta}) \,=\, \sum_{i=1}^{n} (\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}})\, e^{-\boldsymbol{Z}_i' \boldsymbol{\beta}} \, \frac{N_i(C_i)}{\hat{S}_0(C_i^-; \boldsymbol{\beta})^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})}}\,,$$

where

$$\hat{S}_0(t; \boldsymbol{\beta}) \,=\, \exp\left[-\int_0^t \frac{d\,\tilde{N}(s)}{\sum_{i=1}^n I(s \leq C_i)\, e^{\boldsymbol{Z}_i' \boldsymbol{\beta}}}\right]\,.$$

This statistic adjusts for the bias due to the difference in the distributions of the C_i's and represents the adjusted observed treatment differences. Of course, in general, the parameter $\boldsymbol{\beta}$ is unknown. Because we observe complete data on the C_i's, it is then natural to estimate $\boldsymbol{\beta}$ by the partial likelihood estimator, $\hat{\boldsymbol{\beta}}$, defined as the solution to

$$\boldsymbol{U}(\boldsymbol{\beta}) \,=\, \sum_{i=1}^{n} \int_0^\infty \left[\boldsymbol{Z}_i \,-\, \frac{\sum_{j=1}^n I(t \le C_j)\, e^{\boldsymbol{Z}'_i \boldsymbol{\beta}}\, \boldsymbol{Z}_j}{\sum_{j=1}^n I(t \le C_j)\, e^{\boldsymbol{Z}'_i \boldsymbol{\beta}}} \right] d\tilde{N}_i(t) \,=\, 0\,.$$

Therefore, the hypothesis H_0 can be tested using the test statistic $\boldsymbol{U}_{cc}(\hat{\boldsymbol{\beta}})$.

Define $A(\boldsymbol{\beta}) \,=\, \partial \boldsymbol{U}_{cc}(\boldsymbol{\beta})\,/\partial\boldsymbol{\beta}$ and $B(\boldsymbol{\beta}) \,=\, -\,\partial \boldsymbol{U}(\boldsymbol{\beta})\,/\partial\boldsymbol{\beta}$. Also define

$$R(t) \,=\, \sum_{i=1}^n (\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}}) \int_t^\infty \frac{N_i(s)\, d\tilde{N}_i(s)}{\{\hat{S}_0(s, \hat{\boldsymbol{\beta}})\}^{\exp(\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}})}}\,,$$

$$\hat{a}_i \,=\, (\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}})\, e^{-\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}} \int_0^\infty \frac{N_i(t)\, d\tilde{N}_i(t)}{\{\hat{S}_0(t-, \hat{\boldsymbol{\beta}})\}^{\exp(\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}})}}\,,$$

$$\hat{b}_i \,=\, \int_0^\infty \frac{R(t)}{\sum_{i=1}^n I(t \le C_i)\, e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}} \left[d\tilde{N}_i(t) \,-\, \frac{I(t \le C_i)\, e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}}{\sum_{i=1}^n I(t \le C_i)\, e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}}\, d\tilde{N}(t) \right]$$

and

$$\begin{aligned} \hat{\alpha}_i \,=\, & \int_0^\infty \left[\boldsymbol{Z}_i - \frac{\sum_{i=1}^n I(t \le C_i) e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}} \boldsymbol{Z}_i}{\sum_{i=1}^n I(t \le C_i) e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}} \right] \\ & \times \left[d\tilde{N}_i(t) - \frac{I(t \le t_i) e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}}{\sum_{i=1}^n I(t \le C_i) e^{\boldsymbol{Z}'_i \hat{\boldsymbol{\beta}}}}\, d\tilde{N}(t) \right], \end{aligned}$$

$i \,=\, 1, ..., n$. Sun (1999) shows that if n is large and H_0 is true, the distribution of $\boldsymbol{U}_{cc}(\hat{\boldsymbol{\beta}})$ can be approximated by the multivariate normal distribution with mean 0 and covariance matrix $V_{cc} \,=\, H(\hat{\boldsymbol{\beta}})\, \Gamma\, H'(\hat{\boldsymbol{\beta}})$, where $H(\boldsymbol{\beta}) \,=\, (\, I\,,\, A(\boldsymbol{\beta})\, B^{-1}(\boldsymbol{\beta})\,)$, I denotes the $p \times p$ identity matrix, and

$$\Gamma \,=\, \sum_{i=1}^n \begin{pmatrix} \hat{a}_i + \hat{b}_i \\ \hat{\alpha}_i \end{pmatrix} \left(\hat{a}'_i + \hat{b}'_i\,,\, \hat{\alpha}'_i \right)\,.$$

Therefore, a test of the hypothesis H_0 can be based on the statistic $U^*_{cc} \,=\, \boldsymbol{U}_{cc}(\hat{\boldsymbol{\beta}})'\, V_{cc}^{-1}\, \boldsymbol{U}_{cc}(\hat{\boldsymbol{\beta}})$, whose null distribution can be approximated by the χ^2 distribution with p degrees of freedom.

This approach requires that the distribution of the observation times, C_i's, can be described by the PH model, which can be easily checked in practice because one observes complete data on the C_i's. However, it is known that the PH model fits most failure time data reasonably well.

Another way to develop a test for H_0 is to start with the statistic U_{cs} defined in (4.2) and to adjust for the bias introduced by the distributions of the C_i's. In using U_{cs}, the adjustment would involve estimation of the density functions of observation times, which makes the approach complicated and unattractive.

4.3 Rank-based Comparison Procedures

Consider general or case II interval-censored failure time data and suppose that the observed data are

$$\{\,(L_i, R_i], \boldsymbol{Z}_i\,;\, i \,=\, 1, ..., n\,\}$$

from n independent subjects with each receiving one of $p\,+\,1$ different treatments. In the above, $(L_i, R_i]$ denotes the interval within which the survival event of interest is observed to occur for subject i, and $\boldsymbol{Z}_i$ is the p-dimensional vector of treatment indicators. The goal is to test the hypothesis H_0 : the $p+1$ survival functions corresponding with the treatments are identical.

Several approaches can be taken to develop test procedures that make use of the rankings of the underlying unobserved true failure times, T_i, and are similar to the methods developed for complete or right-censored failure time data. In this section, we focus attention on the approach given in Zhao and Sun (2004) that directly generalizes the log-rank test, which is the most commonly used method for right-censored data, because of its simplicity and ease of interpretation and implementation. A brief discussion on some other approaches is followed in Section 4.3.2.

4.3.1 Generalized Log-rank Test

Let $\hat{S}_0$ denote the maximum likelihood estimator of the common survival function $S_0(t) \,=\, Pr(T_i \,>\, t)$ under H_0 and $s_1 \,<\, \cdots \,<\, s_m$ the ordered distinct time points of $\{L_i, R_i\;,\; i \,=\, 1, ..., n\}$ at which $\hat{S}_0$ has jumps. For notational convenience, we assume that there exists a time point $s_{m+1} > s_m$ at which $\hat{S}_0$ has all the remaining mass. That is, $\hat{S}_0(t)$ is equal to zero for $t \,\geq\, s_{m+1}$. For each pair (i, j), define $\alpha_{ij} \,=\, I(\,s_j \in (L_i, R_i]\,)$, the indicator of the event $s_j \in (L_i, R_i]$, $i \,=\, 1, ..., n$, $j \,=\, 1, ..., m+1$. To develop a statistic of log-rank type for H_0, we recall that the log-rank test statistic is the summation of the observed minus expected numbers of deaths or events. Thus we need to determine the number of failures and the number at risk for each observed failure time, which we refer to as the failure and risk numbers.

For this, define $\delta_i \,=\, 0$ if the observation on the failure time T_i for the ith subject is right-censored and 1 otherwise, $i \,=\, 1, ..., n$. That is, $\delta_i \,=\, I(R_i \,\leq\, s_m)$. Also define $\rho_{ij} \,=\, I(\,\delta_i = 0\,,\, L_i \geq s_j\,)$, which is 1 if T_i is right-censored and subject i is still at risk at s_j-, $i \,=\, 1, ..., n$, $j \,=\, 1, ..., m$. Then if H_0 is true and $S_0(t)$ is treated as known and given by $\hat{S}_0(t)$, natural estimates of the overall observed failure and risk numbers at time s_j are given by

$$d_j = \sum_{i=1}^{n} \delta_i \, \frac{\alpha_{ij}[\hat{S}_0(s_j-) \,-\, \hat{S}_0(s_j)]}{\sum_{u=1}^{m+1} \alpha_{iu}[\hat{S}_0(s_u-) \,-\, \hat{S}_0(s_u)]}$$

and

$$n_j = \sum_{r=j}^{m+1} \sum_{i=1}^{n} \delta_i \, \frac{\alpha_{ir} [\hat{S}_0(s_r-) \, - \, \hat{S}_0(s_r)]}{\sum_{u=1}^{m+1} \alpha_{iu} [\hat{S}_0(s_u-) \, - \, \hat{S}_0(s_u)]} \, + \sum_{i=1}^{n} \rho_{ij} \, ,$$

respectively, $j = 1, ..., m$. Similarly estimates of the observed failure and risk numbers at time s_j, $j = 1, ..., m$, for treatment group l, $l = 1, ..., p+1$, are

$$d_{j\,l} \, = \sum_{i}^{l} \delta_i \, \frac{\alpha_{ij} [\hat{S}_0(s_j-) \, - \, \hat{S}_0(s_j)]}{\sum_{u=1}^{m+1} \alpha_{iu} [\hat{S}_0(s_u-) \, - \, \hat{S}_0(s_u)]}$$

and

$$n_{j\,l} \, = \sum_{r=j}^{m+1} \sum_{i}^{l} \delta_i \, \frac{\alpha_{ir} [\hat{S}_0(s_r-) \, - \, \hat{S}_0(s_r)]}{\sum_{u=1}^{m+1} \alpha_{iu} [\hat{S}_0(s_u-) \, - \, \hat{S}_0(s_u)]} \, + \sum_{i}^{l} \rho_{ij} \, ,$$

respectively, where $\sum_i^l$ denotes the summation over all subjects in population l. It is apparent that if right-censored data are available, the above estimates reduce to the observed failure and risk numbers that are used in the construction of the log-rank test statistic.

To test H_0, motivated by the log-rank statistic, one can use the statistic $\boldsymbol{U}_r = (U_{r,1}, ..., U_{r,p+1})'$, where

$$U_{r,l} \, = \sum_{j=1}^{m} \left(d_{j\,l} \, - \, \frac{n_{j\,l} d_j}{n_j} \right) \, ,$$

the summation of the observed numbers of failures minus the expected numbers of failures. To estimate the covariance matrix of $\boldsymbol{U}_r$, Zhao and Sun (2004) give the following multiple imputation approach. Let M be a prespecified integer. For each b $(1 \leq b \leq M)$,

Step 1. For each i, if $\delta_i = 0$, let $T_i^{(b)} = L_i$ and $\delta_i^{(b)} = 0$ and if $\delta_i = 1$, let $T_i^{(b)}$ be a random number drawn from the conditional probability function

$$f_i(s) \, = \, Pr\{ \, T_i^{(b)} \, = \, s \} \, = \, \frac{\hat{S}_0(s-) \, - \, \hat{S}_0(s)}{\hat{S}_0(L_i) \, - \, \hat{S}_0(R_i)}$$

over the s_j's that belong to $(L_i, R_i]$ and $\delta_i^{(b)} = 1$, $i = 1, ..., n$. That is, we generate a set of right-censored data $\{(T_i^{(b)}, \delta_i^{(b)}) \, ; \, i = 1, ..., n\}$ with the same censoring indicators $\delta_i^{(b)}$ for all b.

Step 2. Given $\{(T_i^{(b)}, \delta_i^{(b)}) \, ; \, i = 1, ..., n\}$, determine the corresponding observed failure and risk numbers, i.e., $d_j^{(b)}$'s, $n_j^{(b)}$'s, $d_{j\,l}^{(b)}$'s and $n_{j\,l}^{(b)}$'s, from all subjects and from the subjects in each treatment group, respectively. Let $\boldsymbol{U}^{(b)}$ denote the log-rank statistic based on this set of generated right-censored data, which is defined like $\boldsymbol{U}_r$ with d_j, n_j, $d_{j\,l}$ and $n_{j\,l}$ replaced by the $d_j^{(b)}$, $n_j^{(b)}$, $d_{j\,l}^{(b)}$ and $n_{j\,l}^{(b)}$, respectively. Then calculate the estimate of the covariance matrix of $\boldsymbol{U}^{(b)}$ given by $\hat{\boldsymbol{V}}^{(b)} = \hat{\boldsymbol{V}}_1^{(b)} + ... + \hat{\boldsymbol{V}}_m^{(b)}$, where $\hat{\boldsymbol{V}}_j^{(b)}$ is a $(p+1) \times (p+1)$ matrix with elements

$$(\hat{V}_j^{(b)})_{l\,l} = \frac{n_{j\,l}^{(b)}\,(n_j^{(b)} - n_{j\,l}^{(b)})\,d_j^{(b)}\,(n_j^{(b)} - d_j^{(b)})}{(n_j^{(b)})^2\,(n_j^{(b)} - 1)}\,,\; l = 1, ..., p+1\,,$$

and

$$(\hat{V}_j^{(b)})_{l_1 l_2} = -\,\frac{n_{j\,l_1}^{(b)}\,n_{j\,l_2}^{(b)}\,d_j^{(b)}\,(n_j^{(b)} - d_j^{(b)})}{(n_j^{(b)})^2\,(n_j^{(b)} - 1)}\,,\; 1 \le l_1 \ne l_2 \le p+1\,,$$

for $j = 1, ..., m$. Here $\hat{\boldsymbol{V}}^{(b)}$ is obtained by using a conditional multivariate hypergeometric distribution at each time point s_j, $j = 1, ..., m$.
Step 3. Repeat the steps 1 to 2 for each $b = 1, ..., M$, and then the covariance matrix of $\boldsymbol{U}_r$ can be estimated by $\hat{\boldsymbol{V}}_r = \hat{\boldsymbol{V}}_{r,1} + \hat{\boldsymbol{V}}_{r,2}$, where

$$\hat{\boldsymbol{V}}_{r,1} = \frac{1}{M} \sum_{b=1}^{M} \hat{\boldsymbol{V}}^{(b)}\,,$$

$$\hat{\boldsymbol{V}}_{r,2} = \left(1 + \frac{1}{M}\right) \frac{\sum_{b=1}^{M} [\boldsymbol{U}^{(b)} - \bar{\boldsymbol{U}}_r]\,[\boldsymbol{U}^{(b)} - \bar{\boldsymbol{U}}_r]'}{M - 1}$$

with $\bar{\boldsymbol{U}}_r = \sum_{b=1}^{M} \boldsymbol{U}^{(b)}\,/M$.

With $\hat{\boldsymbol{V}}_r^-$ denoting a generalized inverse of $\hat{\boldsymbol{V}}_r$, the hypothesis H_0 can be tested using the statistic $U_r^* = \boldsymbol{U}_r'\,\hat{\boldsymbol{V}}_r^-\,\boldsymbol{U}_r$, whose null distribution can be approximated by the χ^2 distribution with p degrees of freedom. A similar approach, which is discussed in Pan (2000b), uses the imputed statistic $\bar{\boldsymbol{U}}_r$ instead of $\boldsymbol{U}_r$. However, $\boldsymbol{U}_r$ has a better interpretation than $\bar{\boldsymbol{U}}_r$. Sun (1996) discussed another procedure in which the test statistic is better motivated for interval-censored data, but the procedure given there does not reduce to the log-rank test for right-censored data. It also can be invalid if the percentage of purely right-censored observations is large.

4.3.2 Discussion

For the two-sample problem ($p = 1$), $U_{r,1}$, the first component of $\boldsymbol{U}_r$ given in the previous subsection can be written as

$$U_{r,1} = \sum_{j=1}^{m} \frac{n_{j\,1} n_{j\,2}}{n_{j\,1} + n_{j\,2}} \left(\frac{d_{j\,1}}{n_{j\,1}} - \frac{d_{j\,2}}{n_{j\,2}} \right) = \int_0^\infty \frac{Y_1(t) Y_2(t)}{Y_1(t) + Y_2(t)} \left[d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t) \right].$$

Here $Y_l(t) = \sum_{j : s_j \le t} n_{j\,l}$ and

$$\hat{\Lambda}_l(t) \,=\, \sum_{j\,:\,s_j\,\leq\,t} \frac{d_{j\,l}}{n_{j\,l}}$$

is an estimator of the cumulative hazard function corresponding with treatment l, $l \,=\, 1,\, 2$. That is, the statistic $\boldsymbol{U}_r$ is an integrated weighted difference between estimated cumulative hazard functions. For the case of $p \,=\, 1$, we can generalize the above test by employing the following weighted statistics

$$\int_0^\infty W(t)\,\frac{Y_1(t)\,Y_2(t)}{Y_1(t)\,+\,Y_2(t)}\,\left[\,d\,\hat{\Lambda}_1(t)\,-\,d\,\hat{\Lambda}_2(t)\right]\,, \tag{4.5}$$

where $W(t)$ is a weight process that can depend on observed data. In the case of right-censored data, the statistics in (4.5) give the weighted log-rank test statistics and in this situation, a commonly used class of weight processes is

$$W(t) \,=\, [\,\hat{S}_0(t-)]^\rho\,[\,1\,-\,\hat{S}_0(t-)\,]^\gamma\,,$$

where ρ and γ are non-negative constants.

In general, in order to obtain an efficient test, the selection of an appropriate weight process depends on knowledge about possible alternatives. Consider the class of weight processes given above. For possibly early hazard differences, for example, one may prefer $\rho \,=\, 1$ and $\gamma \,=\, 0$, which gives a decreasing weight process. On the other hand, if there exist possible late hazard differences, one may use $\rho \,=\, 0$ and $\gamma \,=\, 1$, which gives an increasing weight process. For more discussion about this weight process and its selection for right-censored data, see Fleming and Harrington (1991). Most of the selection criteria discussed there apply to interval-censored data although some care is needed.

Another approach for developing a rank-based comparison procedure for H_0 is to follow the linear rank test theory for complete data. Suppose that the survival times, T_i, can be described by the linear model

$$h(T_i) \,=\, \boldsymbol{Z}_i'\,\boldsymbol{\beta} \,+\, \epsilon_i\,.$$

Here h is a strictly increasing function, $\boldsymbol{\beta}$ is the vector of regression parameters representing treatment differences, and ϵ_i is random error with $E(\epsilon_i) \,=\, 0$ and the probability density and distribution functions f_ϵ and F_ϵ, respectively. Then the hypothesis H_0 is equivalent to $\beta \,=\, 0$.

Assume that complete data are observed. The marginal likelihood function of the rank statistic of the T_i's has the form

$$P(r) \,=\, \int \cdots \int_{\tau_{(1)}<\cdots<\tau_{(n)}} \prod_{i=1}^n f_\epsilon(\tau_{(i)}\,-\,\boldsymbol{Z}_{(i)}'\,\boldsymbol{\beta})\,d\tau_{(i)}\,,$$

where $\tau_{(1)} \,<\, \cdots \,<\, \tau_{(n)}$ are the order statistics of the residuals $\{\,h(T_i)\,-\,\boldsymbol{Z}_i'\,\boldsymbol{\beta}\,\}$ and the $\boldsymbol{Z}_{(i)}$'s are the corresponding ordered treatment indicators.

Then a locally most powerful linear rank test for $\boldsymbol{\beta} = 0$ is given by the score statistic $d \log P(r)/\, d\boldsymbol{\beta}$ at $\boldsymbol{\beta} = 0$, which has the form $\sum_{i=1}^{n} c_i\, \boldsymbol{Z}_{(i)}$, where

$$c_i = \int \cdots \int_{\tau_{(1)} < \cdots < \tau_{(n)}} \left[\frac{-d \log f_\epsilon(\tau_{(i)})}{d\tau_{(i)}} \prod_{j=1}^{n} f_\epsilon(\tau_{(j)})\, d\tau_{(j)} \right] .$$

In theory, this approach can be generalized to censored data. However, unlike the case of complete data, it does not usually yield simple rank statistics as shown below. Kalbfleisch and Prentice (2002) discuss this approach for right-censored data.

To generalize this idea to interval-censored data, one may consider the underlying rank vector corresponding to the complete data that are not observed due to interval censoring. That is, define R to be the set of all possible underlying rank vectors that are consistent with the observed interval-censored failure time data and treat it as the observed data rank. This gives the marginal likelihood function

$$\sum_{r \in R} \int \cdots \int_{\tau_{(1)} < \cdots < \tau_{(n)}} \prod_{i=1}^{n} f_\epsilon(\tau_{(i)} \, - \, \boldsymbol{Z}'_{(i)}\, \boldsymbol{\beta})\, d\tau_{(i)} \tag{4.6}$$

and a rank test statistic can be derived as the derivative of the logarithm of this likelihood at $\boldsymbol{\beta} = 0$. However, computation of the statistic and its covariance matrix, which can be obtained by using either the observed Fisher information or permutation approach, is quite complicated compared with the effort needed for the generalized log-rank test discussed in Section 4.3.1. Self and Grossman (1986) investigated this method for the case where $h(t) = \log(t)$.

We can also generalize the linear rank test for complete data by considering the full likelihood function

$$\prod_{i=1}^{n} \int_{h(L_i) - \boldsymbol{Z}'_i \boldsymbol{\beta}}^{h(R_i) - \boldsymbol{Z}'_i \boldsymbol{\beta}} f_\epsilon(\epsilon_i)\, d\epsilon_i \, = \, \prod_{i=1}^{n} [\, F_\epsilon(\, h(R_i) - \boldsymbol{Z}'_i \boldsymbol{\beta}\,) \, - \, F_\epsilon(\, h(L_i) - \boldsymbol{Z}'_i \boldsymbol{\beta}\,)\,]$$

and basing the test of $\boldsymbol{\beta} = 0$ on the resulting score statistic at $\boldsymbol{\beta} = 0$. As with the test based on the marginal likelihood (4.6), this method is very involved computationally. This procedure is discussed further in Fay (1996).

Finally, one also can test H_0 with the rank test discussed in Mantel (1967), which applies to general censored failure time data and ranks observed intervals. The test is a permutation-type approach and could be much less efficient than the methods discussed above, especially when the percentage of overlapping intervals is high.

An important feature of rank-based tests is their invariance under monotone increasing transformations on the survival time. An example of this is the log-rank test, which is well-known to be the locally optimal test within the class of the PH model. In other words, the rank-based tests are in general expected to be efficient for, and sensitive to, alternatives with ordered hazard functions, such as $H_1 : \lambda_2(t) \geq \lambda_1(t)$ for all t in the case of two-sample comparison. Here λ_1 and λ_2 are the hazard functions corresponding to the two treatments involved. However, they may not be sensitive to, or even consistent for, alternatives with ordered survival functions, such as $H_2 : S_2(t) \geq S_1(t)$ for all t. Here S_1 and S_2 are the survival functions corresponding with λ_1 and λ_2, respectively. It should be noted that H_2 does not imply H_1. Methods that are more appropriate for H_2 are discussed in the next section.

4.4 Survival-based Comparison Procedures

Consider the two-sample comparison problem and let S_1 and S_2 denote the survival functions representing the two treatments. To test the hypothesis $H_0 : S_1(t) = S_2(t)$, a class of test statistics that are parallel to the weighted rank-based test statistics given in (4.5) consists of the integrated weighted survival differences,

$$\int_0^\tau W(t)\,[\,\hat{S}_1(t) - \hat{S}_2(t)\,]\,dt\,. \tag{4.7}$$

In (4.7), τ is the largest observation time, $W(t)$ is a weight process that can depend on observed data, and $\hat{S}_1$ and $\hat{S}_2$ are the NPMLEs of S_1 and S_2 based on the separate samples, respectively. The statistics given in both (4.5) and (4.7) have been studied extensively in the case of right-censored data. While the statistics in (4.5) measure the observed rank or estimated hazard differences between the two treatment groups, the statistics in (4.7) measure the estimated survival differences between the two groups. By letting $W(t) = 1$, the statistic in (4.7) gives the difference of the estimated means.

In the following, we consider three methods for testing H_0 that use the statistic defined in (4.7) or a variant. The first applies to situations where the underlying survival variable of interest is discrete and is originally given in Petroni and Wolfe (1994). The other two concern continuous survival variable and are first considered by Fang et al. (2002) and Zhang et al. (2001), respectively. As in Section 4.3, suppose that one observes general interval-censored data $\{\,(L_i, R_i], Z_i\,;\, i = 1, ..., n\,\}$ and that the distributions of the L_i and R_i's are identical across the two treatment groups.

4.4.1 Comparison with Discrete Survival Time

Suppose that the underlying survival times, T_i, take only discrete values $0 < s_1 < ... < s_m < s_{m+1} = \infty$. Then under H_0, the likelihood function is proportional to

$$L(\mathbf{p}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij}\, p_j \right) ,$$

where $\alpha_{ij} = I(s_j \in (L_i, R_i])$ and $\mathbf{p} = (p_1, ..., p_m)'$ with $p_j = Pr(T_i = s_j \,|\, H_0)$ and $\sum_{j=1}^m p_j = 1$.

Assume that the weight process in (4.7) is a step process with jumps only at the s_j's and converges in probability to a deterministic function. Let $\hat{\mathbf{p}}$ denote the maximum likelihood estimator of $\mathbf{p}$ obtained from the likelihood function $L(\mathbf{p})$ and $I(\hat{\mathbf{p}})$ the observed Fisher information matrix from $L(\mathbf{p})$ at $\mathbf{p} = \hat{\mathbf{p}}$. Taking $\tau = s_m$, the test statistic in (4.7) then has the form

$$U_d = \sum_{j=1}^{m-1} w_j \left[\hat{S}_1(s_j) - \hat{S}_2(s_j) \right] (s_{j+1} - s_j) \,,$$

the summation over all possible failure time points of the weighted differences between the two estimated survival probabilities, where $w_j = W(s_j)$, $j = 1, ..., m-1$. Assume that the numbers of subjects within both treatment groups go to infinity as n goes to infinity. Then using maximum likelihood theory, Petroni and Wolfe (1994) show that under H_0 with large n, the distribution of U_d can be approximated by the normal distribution with mean zero and variance

$$\hat{\sigma}_d^2 = \sum_{j_1=1}^{m-1} \sum_{j_2=1}^{m-1} w_{j_1} w_{j_2} (s_{j_1+1} - s_{j_1}) (s_{j_2+1} - s_{j_2}) \sum_{k_1=1}^{j_1} \sum_{k_2=1}^{j_2} a_{k_1 k_2} \,,$$

where $a_{k_1 k_2}$ is the element of the inverse of $I(\hat{\mathbf{p}})$ at the k_1th row and k_2th column.

Thus the hypothesis H_0 can be tested by employing the statistic $U_d^* = U_d / \hat{\sigma}_d$ with standard normal critical values. As discussed in Section 4.3.2, the selection of an appropriate weight process W is usually important and depends on alternatives of interest. The suggestions given for various types of alternatives and right-censored data generally apply to interval-censored data. Some discussion on this matter for the discrete survival time data can be found in Petroni and Wolfe (1994).

4.4.2 Comparison I with Continuous Survival Time

In this subsection, we discuss another special case of the test statistics given in (4.7) for testing H_0. Assume that the T_i's are continuous variables and for convenience, assume that the support of the survival functions belongs to $[0, \tau]$. Also assume that n_1 of the n individuals come from the population with survival function S_1 and the remaining n_2 individuals come from the population with survival function S_2, where $n = n_1 + n_2$. To present the method, it is convenient to formulate the observed data as

$$\{ \, (u_i, v_i, \delta_{1i}, \delta_{2i}, \delta_{3i}) \, ; \, i = 1, \cdots, n \, \} \,,$$

where u_i and v_i are observed values of the two random variables U_i and V_i with $U_i \leq V_i$ and the δ_{li}'s are the observed values of the indicator variables $\Delta_{1i} =$

$I(T_i \le U_i)$, $\Delta_{2i} = I(U_i < T_i \le V_i)$ and $\Delta_{3i} = 1 - \Delta_{1i} - \Delta_{2i}$. Let $H(u,v)$ denote the joint cumulative distribution function of U and V and H_1 and H_2 the marginal cumulative distribution functions of U and V, respectively. Also let $h(u,v)$, $h_1(u)$, and $h_2(v)$ denote the corresponding density functions. For the statistics given in (4.7), Fang et al. (2002) give the following result.

Assume that $n_1/n \to p$ as $n \to \infty$, where $0 < p < 1$. Also assume that $W(t) = w(t)$ is a deterministic function with a bounded derivative on $[0,\tau]$. Then under H_0 and some regularity conditions, as $n \to \infty$, the statistic

$$U_c = \sqrt{\frac{n_1\, n_2}{n}} \int_0^{\tau} w(t)\, [\, \hat{S}_1(t) - \hat{S}_2(t)\,]\, dt$$

has an asymptotic normal distribution with mean 0. One key regularity condition is condition (I3) given in Section 3.6.2 and requires that U_i and V_i should not be too close together. That is, it is assumed that no exact failure times are observed.

For consistent estimation of the asymptotic variance of the statistic U_c, as before, let $\hat{S}_0$ denote the NPMLE of the common survival function under H_0, $\hat{F}_0(t) = 1 - \hat{S}_0(t)$ and $0 < s_1 < ... < s_m < \tau$ denote the time points at which $\hat{S}_0$ and $\hat{F}_0$ have jumps. Also let $a_j = \hat{F}_0(s_j)$, $j = 1, ..., m$, and $\phi_{w,\hat{F}_0}$ denote the solution to the following Fredholm integral equation

$$\phi_{w,\hat{F}_0}(t) = d_{\hat{F}_0}(t) \left[w(t) - \int_0^{\tau} \frac{\phi_{w,\hat{F}_0}(t) - \phi_{w,\hat{F}_0}(t')}{|\hat{F}_0(t) - \hat{F}_0(t')|}\, h^*(t',t)\, dt' \right] ,$$

where

$$d_{\hat{F}_0}(t) = \frac{\hat{F}_0(t)\, [1 - \hat{F}_0(t)]}{h_1(t)\, [1 - \hat{F}_0(t)] + h_2(t)\, \hat{F}_0(t)}$$

and $h^*(t',t) = h(t',t) + h(t,t')$. Then $\phi_{w,\hat{F}_0}$ is absolutely continuous with respect to $\hat{F}_0$ and a step function with jumps at the s_j's.

Let $\hat{H}$, $\hat{H}_1$ and $\hat{H}_2$ denote the empirical distributions of (U,V), U and V, respectively. Define $y_j = \phi_{w,\hat{F}_0}(s_j)$,

$$\Delta_j(h_r) = \int_{s_j}^{s_{j+1}} h_r(t)\, dt \approx \int_{s_j}^{s_{j+1}} d\hat{H}_r(t) ,$$

$$\Delta_{jl}(h) = \int_{u=s_j}^{s_{j+1}} \int_{s_l}^{s_{l+1}} h(u,v)\, dvdu \approx \int_{s_j}^{s_{j+1}} \int_{s_l}^{s_{l+1}} d\hat{H}_0(u,v)$$

and

$$d_j = \frac{a_j\, (1 - a_j)}{\Delta_j(h_1)(1 - a_j) + \Delta_j(h_2)\, a_j} ,$$

$j\, ,\, l = 1, \cdots, m$, $r = 1, 2$. Then it can be shown that the vector $\mathbf{y} = (y_1, \cdots, y_m)'$ is the unique solution to the following set of linear equations

$$y_j \left[d_j^{-1} + \sum_{l<j} \frac{\Delta_{lj}(h)}{a_j - a_l} + \sum_{l>j} \frac{\Delta_{jl}(h)}{a_l - a_j} \right] = \Delta_j(w) + \sum_{l<j} \frac{\Delta_{lj}(h)}{a_j - a_l} y_l + \sum_{l>j} \frac{\Delta_{jl}(h)}{a_l - a_j} y_l$$

for $j = 1, 2, ..., m$.

Furthermore, define

$$\tilde{\theta}_{w,\hat{F}_0}(u, v, \delta_1, \delta_2) = -\delta_1 \frac{\phi_{w,\hat{F}_0}(u)}{\hat{F}_0(u)} - \delta_2 \frac{\phi_{w,\hat{F}_0}(v) - \phi_{w,\hat{F}_0}(u)}{\hat{F}_0(v) - \hat{F}_0(u)} + \delta_3 \frac{\phi_{w,\hat{F}_0}(v)}{1 - \hat{F}_0(v)}$$

and for $(\delta_1, \delta_2) = (0, 0)$, $(0, 1)$, $(1, 0)$,

$$\hat{Q}(u, v, \delta_1, \delta_2) = \sum_i \Delta\hat{H}(u_i, v_i) \hat{F}_0^{\delta_{1i}}(u_i) \{\hat{F}_0(v_i) - \hat{F}_0(u_i)\}^{\delta_{2i}} \{1 - \hat{F}_0(v_i)\}^{1-\delta_{1i}-\delta_{2i}},$$

which is the empirical distribution of the vector $(U_i, V_i, \Delta_{1i}, \Delta_{2i})$, where the summation is over $\{\, i\,;\,,\, u_i \leq u, v_i \leq v, \delta_{1i} = \delta_1, \delta_{2i} = \delta_2 \,\}$. Then Fang et al. (2002) show that a consistent estimator of the asymptotic variance of U_c is given by

$$\|\tilde{\theta}_{w,\hat{F}_0}\|^2 = \int \tilde{\theta}^2_{w,\hat{F}_0}(u, v, \delta_1, \delta_2)\, d\hat{Q}(u, v, \delta_1, \delta_2)\,.$$

Hence the hypothesis H_0 can be tested using the statistic $U_c / \, \|\tilde{\theta}_{w,\hat{F}_n}\|$ with standard normal critical values.

For the weight function $w(t)$, the simplest choices include $w(t) = 1$, $w(t) = 1/(t+1)$ (decreasing) and $w(t) = 1 - 1/(t+1)$ (increasing). The discussion and remarks given in Sections 4.3.2. and 4.4.1 concerning the weight function apply here.

4.4.3 Comparison II with Continuous Survival Time

Consider the same situation as in the last subsection and suppose that the same assumptions hold. Using the same notation, first we notice that the test statistic U_c is motivated partly by the functional $\int_0^\tau w(t)\, S(t)\, dt$ and as discussed in Section 3.6, $\hat{S}_l$ $(l = 1, 2)$ converges at the rate of $n^{-1/3}$, while its functional converges at the rate of $n^{-1/2}$. Also, recall that the NPMLE $\hat{S}_0$ has a closed form in the case of current status data but does not for general interval-censored data. This plus the complexity of the variance estimator of U_c suggests that one might develop an alternative to U_c for testing H_0 by considering the functional given above and the two separate sets of current status data

$$\{\, (u_i, \delta_{1i})\,;\, i = 1, \cdots, n \,\} \quad and \quad \{\, (v_i, 1 - \delta_{3i})\,;\, i = 1, \cdots, n \,\} \tag{4.8}$$

on the (U_i, Δ_{1i})'s and $(V_i, 1 - \Delta_{3i})$'s, respectively.

Under H_0, let $\hat{S}_{0p}$ denote the NPMLE of the common survival function, S_0, that is based on the current status data obtained by combining the two sets of current status data in (4.8) and treating them as independent data sets. Also, let $\hat{S}_{lp}$ denote the estimator of S_l that is defined like $\hat{S}_{0p}$, but is based only on observed data corresponding with S_l, $l = 1, 2$. Motivated by the functional

$$\int\int [\, w(u)\, S_l(u) \; + \; w(v)\, S_l(v)\,]\, d\, H(u,v) \; ,$$

the sum of the two functionals corresponding with the two sets of current status data in (4.8), similar to U_c, we construct the test statistic

$$U_c^* = \sqrt{n} \int_0^\tau \left[w(u)\{\hat{S}_{1p}(u) - \hat{S}_{2p}(u)\} d\hat{H}_1(u) + w(v)\{\hat{S}_{1p}(v) - \hat{S}_{2p}(v)\} d\hat{H}_2(v) \right] .$$

Zhang et al. (2001) show that under some regularity conditions and the hypothesis H_0, the distribution of U_c^* can be approximated by the normal distribution with mean zero and variance

$$\begin{aligned} \hat{\sigma}_c^{*2} \; = \; & \frac{n^2}{n_1 n_2} \int_0^\tau \left[w^2(u)\hat{S}_{0p}(u)\{1 - \hat{S}_{0p}(u)\} d\hat{H}_1(u) \right. \\ & \left. + \, w^2(v)\hat{S}_{0p}(v)\{1 - \hat{S}_{0p}(v)\} d\hat{H}_2(v) \right] \\ + \, & \frac{2n^2}{n_1 n_2} \int\int_{0 \le u \le v \le \tau} \{1 - \hat{S}_{0p}(u)\}\hat{S}_{op}(v) w(u) w(v) d\hat{H}(u,v) \end{aligned}$$

if both n_1 and n_2 are large. Hence the hypothesis H_0 can be tested by using the statistic $U_c^* / \, \hat{\sigma}_c^*$ and the standard normal critical values.

4.4.4 Discussion

The method based on the statistics defined in (4.7) can be easily generalized to the $p+1$ sample comparison problem. For the situation, the statistics can be replaced by

$$\int_0^\tau W_l(t)\, [\, \hat{S}_l(t) \; - \; \hat{S}_0(t)\,]\, dt \; .$$

In the latter, W_1, ..., W_{p+1} are weight processes as before, $\hat{S}_l$ denotes the NPMLE of the survival function corresponding with sample l, $l = 1, ..., p+1$, and $\hat{S}_0$ is the NPMLE of the common survival function under the hypothesis of equality.

There exist several differences between the two test procedures discussed in Sections 4.4.2 and 4.4.3. It is apparent that a major advantage of the method based on U_c over that based on U_c^* is that the former is more efficient because it fully makes use of the observed interval-censored data. In contrast, the latter divides the observed information on the same subject into two parts and treats them as independent samples. The advantage of the latter is its simplicity because all quantities involved, including the variance estimator, have closed forms. Fang et al. (2002) give an alternative approach for estimating the variance of U_c that is based on a simple bootstrap procedure. Computationally, the variance estimation method is much simpler, but no theoretical justification is provided.

4.5 Examples

Three illustrative examples are discussed in this section. The first two examples concern current status data from tumorigenicity experiments and the third example considers the analysis of the breast cancer data discussed in Sections 1.2.2, 3.4.4, and 3.5.2. In the applications of the methods presented in Section 4.2 for current status data, we focus on the situation where the distribution of the observation times C_i's may depend on treatment indicators.

4.5.1 Analysis of Tumorigenicity Experiments

We first consider the lung tumor data discussed in Section 1.2.1. As mentioned before, lung tumors are usually regarded as nonlethal, and thus one can reasonably assume that the data are current status data with respect to the time to tumor onset. Also the death times, which serve as observation times, can be assumed to be independent of tumor onset times within each treatment group. For the comparison of the rates of development of nonlethal tumors, traditional methods can be classified into two types: interval-based tests and model-based tests. In interval-based tests, animals are grouped into several time intervals according to age at death, and the numbers of animals who die with and without tumors within each interval are counted and used for the comparison. It is apparent that the analysis results can vary widely depending on the choice of intervals. The model-based tests refer to approaches developed by specifying certain models for tumor prevalence or risk and for these tests, it may be difficult to justify the assumed model in practice. In contrast, the methods given in Section 4.2 do not depend on either the choice of intervals or the assumed model.

To compare the tumor incidence rates between the two treatment groups, we first note that for the data in Table 1.3, the numbers of animals who had developed tumors at their deaths are 27 and 35 in the conventional and germ-free environments, respectively. This gives empirical tumor development rates of 0.28 and 0.73 without considering death time information and suggests that there is a difference between the tumor rates in the two groups. To make a statistical comparison, we need to determine if the distribution of the death times depends on the treatment. Figure 4.1 presents separate Kaplan-Meier estimators of the survival functions of the death times, C_i, for animals in the two treatment groups. It seems that the distributions are different and the mice in the germ-free environment had significantly longer survival time than

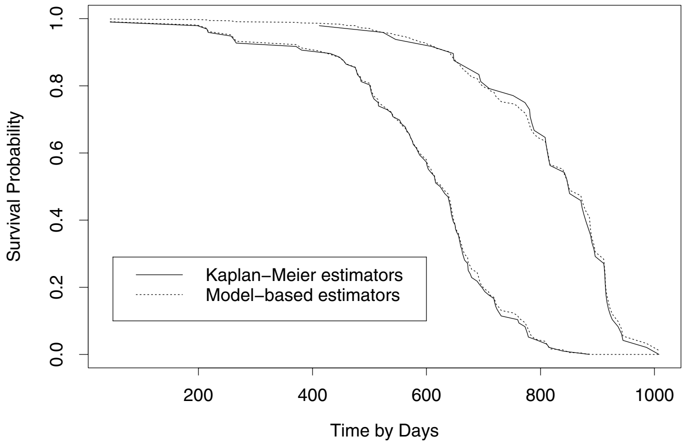

Fig. 4.1. Estimates of survival functions of death times: top, GE; bottom, CE.

those in the conventional environment. This suggests the use of the approach discussed in Section 4.2.2.

Define $Z_i = 0$ for the animals in the conventional environment and 1 otherwise. Fitting the PH model (4.4) yields $\hat{\beta} = -1.9627$ and the test of $\beta = 0$ gives a p-value of less than 0.0001, which further indicates that the distributions of the observation times differ. Application of the method given in Section 4.2.2 results in $U_{c\,c}^{*} = 4.852$ with a p-value of 0.028. The result indicates that there is a significant difference between the lung tumor incidence rates of the mice in the two treatments groups. The mice in the germ-free environment seem to have higher tumor incidence rate than those in the conventional environment. For comparison, we also apply the procedure based on the statistic $U_{c\,w}$ given in Section 4.2.1 and obtain $U_{c\,w} = 1.194$ and a p-value of 0.0009 for testing no difference between the tumor incidence rates. Note that although both approaches give similar results, like the empirical comparison approach, the method that ignores the difference between the distributions of the observation times overestimates the treatment difference. An explanation for this is that by forcing the same death rate on the two groups, the death rate difference is added to the tumor incidence rate difference.

To check the validity of model (4.4) and thus the above result, we obtain estimates of the survival functions of the death times for animals in the two treatment groups under the model (4.4). The estimates are included in Figure 4.1 for comparison. It can be seen from that figure that both estimates are quite close to the corresponding Kaplan-Meier estimators, and this suggests that the model is reasonable for this set of data.

Table 4.1. Survival times in weeks for 100 male F344 rats with testicular tumors

Group	Tumor presence	Survival times in weeks (C_i)
Control	With tumor ($\delta_i = 1$)	73, 79, 87, 91, 95, 96(2), 98, 100(2), 101, 103 104, 108(36)
	No tumor ($\delta_i = 0$)	23
Treatment	With tumor ($\delta_i = 1$)	60, 68, 70, 73, 74, 78, 79, 82(2), 84, 90(3), 92 96(2), 100, 103(2), 104, 105, 106, 108(7)
	No tumor ($\delta_i = 0$)	2, 3, 5, 8(3), 9, 10, 12(2), 14, 24(2), 26, 38 40, 42, 47, 52, 55, 108

As another example, consider the data given in Table 4.1 from a tumorigenicity study on 100 male F344 rats about testicular tumors, which are also known to be relatively nonlethal. The data are reproduced from Lagakos and Louis (1988) and consist of survival times by weeks of the rats and the status of tumor presence or absence at their death times. The numbers in parentheses represent the numbers of the rats with the same survival time. The study involves two groups, control (50 rats, $z_i = 0$) and treatment (50 rats, $z_i = 1$) groups, in which the rats were exposed by gavage to 0 or 60mg%/kg of commercial grade toluene diisocyanate, respectively. It can be seen from the table that in the control group, only one animal died without a tumor. The observed tumor rates are 98% and 58% for the animals in the control and treatment groups, respectively. This suggests that there may exist a difference between the tumor incidence rates of the rats in the two groups with the rats in the control group having a higher rate.

Also it can be seen from Table 4.1 that in the control group, 36 animals survived up to 108 weeks, the time at which the rats were sacrificed. In contrast, only 8 animals survived up to 108 weeks in the treatment group. That is, the rats in the control group seem to survive much longer than those in the treatment group. As with the lung tumor data, under model (4.4), we test for equality of the survival functions of the animals natural death times of the two groups and obtain a p-value of less than 0.0001. This indicates that the approach in Section 4.2.2 should be applied to take into account the difference in death times in the comparison of the testicular tumor incidence rates.

Applying the method given in Section 4.2.2 to this set of testicular tumor data, we obtain $U_{cc}(\hat{\beta}) = 4.376$ and a p-value of 0.054 for testing equality of the tumor incidence rates in the two groups. The result suggests that the rats in the treatment group had a moderately higher tumor incidence rate than those in the control group. Note that unlike the first example, this is quite different from the simple empirical comparison of tumor rates (98% against 58%), which indicates that the rats in the control group had a higher tumor rate. For comparison, again we apply the statistic U_{cw} to the data and obtain $U_{cw} = -1.000$ and a p-value of 0.024 for the comparison of the tumor incidence rates between the two groups. These results suggest that

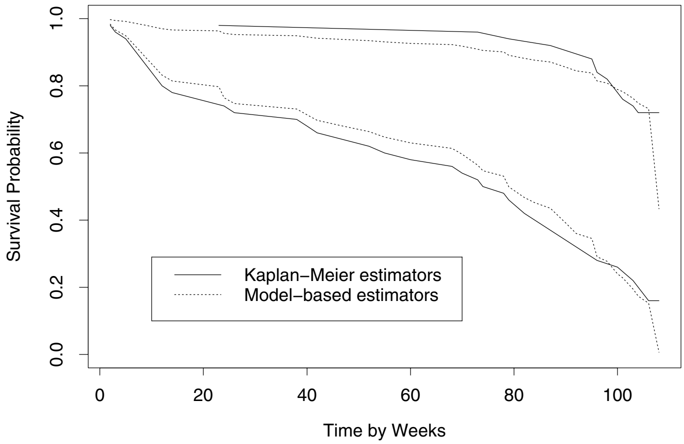

Fig. 4.2. Estimates of survival functions of natural death times: top, control; bottom, treatment.

comparisons, which do not take into account the dependence of the death time distribution on the treatment, can not only overestimate the tumor rate difference, but also can give a misleading direction. In summary, these results indicate that if all the rats had similar survival times, we would see more testicular tumors in the rats belonging to the exposure group.

As with the first example, to check the validity of model (4.4) and hence the results above, we obtain the separate Kaplan-Meier estimators of the survival functions of natural death times of the rats in the two groups. They are displayed in Figure 4.2 along with the corresponding estimators given under model (4.4). The figure indicates that the PH model, (4.4), fits the death times well for this problem.

4.5.2 Analysis of Breast Cancer Data

For the breast cancer data discussed in Sections 1.2.2, 3.4.4, and 3.5.2, our main interest is to compare the early breast cancer patients who were treated with radiation therapy alone to those treated with radiation therapy plus adjuvant chemotherapy with respect to the time to breast retraction. To the medical investigator, it seems that the time between visits or visit times were independent of the times to breast retraction. Also, the assignment of the treatment was not related to this cosmetic result (Finkelstein, 1986). Thus it seems appropriate to apply the methods described in Sections 4.3 and 4.4 to this set of interval-censored data.

For the treatment comparison, we first apply the generalized log-rank test given in Section 4.3.1 and obtain $U_{r,1} = -9.4290$, the component corresponding with the RT group, with an estimated standard error of 3.4185. This gives a p-value of 0.0058 for the comparison and suggests that the patients in the RT group have lower breast retraction rate than those in the RCT group. Next, given the discrete nature of the data, we use the survival-based procedure for discrete failure time data discussed in Section 4.4.1. Assuming that the failure, i.e., breast retraction, can occur only at six-month time points, we obtain $U_d = 8.9697$ and $\hat{\sigma}_d^2 = 5.5652$ with $w(t) = 1$, yielding a p-value of close to 0.0001. This result again indicates that the use of the adjuvant chemotherapy significantly increased the breast retraction risk compared with the use of only radiation therapy. We tried several other discretizing schemes and the weight functions $w(t) = 1/(t+1)$ and $w(t) = 1 - 1/(1+t)$ and got similar p-values. To confirm these results, we further apply the method discussed in Section 4.4.2 with $w(t) = 1$ to the data and obtain $U_c = 42.7130$ with estimated standard deviation $\| \tilde{\theta}_{w,\hat{F}_0} \| = 12.4062$. This corresponds with a p-value of 0.0006 for the comparison of the two treatments and suggests the same conclusion as above.

It is seen that both rank- and survival-based methods give similar results and indicate that the adjuvant chemotherapy significantly increases breast retraction rate. On the other hand, the p-values given and thus the levels of differences suggested by the two types of approaches are quite different. One possible explanation for this is the survival difference between the patients in the two treatment groups is more significant than the corresponding hazard difference.

4.6 Bibliography, Discussion, and Remarks

As with right-censored failure time data, most nonparametric test procedures for interval-censored data can be classified into two categories: rank-based ones and survival-based ones. Research on the ideas behind these two types of test procedures goes back a long way (Kaplan and Meier, 1958). For interval-censored data, some early rank-based and survival-based approaches were given by Mantel (1967) and Peto and Peto (1972), respectively. Following them, many authors studied the nonparametric comparison problem for interval-censored data. As discussed above, for example, Fay (1996), Pan (2000b), Self and Grossman (1986), Sun (1996), and Zhao and Sun (2004) developed some rank-based nonparametric approaches. For the survival-based nonparametric approach, references include Andersen and Ronn (1995), Fang et al. (2002), Petroni and Wolfe (1994), Sun (1999), Sun and Kalbfleisch (1993, 1996), Tang et al. (1995), and Zhang et al. (2001).

Others that also investigated the nonparametric comparison for interval-censored failure time data include Dinse (1994), Fay (1999a), Fay and Shih (1998), Lim and Sun (2003), Pan (1999), Sun, Zhao and Zhao (2005), and

Zhang et al. (2003). In particular, Fay (1999a) and Pan (1999) considered the comparison of several existing procedures, and Lim and Sun (2003) presented a general class of test statistics that include most of existing test statistics as special cases. Sun, Zhao and Zhao (2005) generalized the approach investigated by Peto and Peto (1972). They proposed a class of test statistics for the comparison of $p + 1$ survival functions given by

$$U_\xi = \sum_{i=1}^{n} \boldsymbol{Z}_i \, \frac{\xi\{\hat{S}_0(L_i)\} - \xi\{\hat{S}_0(R_i)\}}{\hat{S}_0(L_i) - \hat{S}_0(R_i)} \,,$$

where ξ is a known function over $(0, 1)$ and the other notation is the same as before. A similar class of statistics is discussed in Section 6.4.3.

For the procedures discussed in Section 4.2, one can add a weight function or weight process like those discussed in Sections 4.3 and 4.4. For example, the statistic U_{cs} given in (4.2) can be generalized to

$$\int_0^\tau W(t)\,[\,\hat{S}_2(t) - \hat{S}_1(t)\,]\,d\,\hat{G}_n(t) \,.$$

Of course, selection of the weight process now becomes an issue that needs thorough investigation, and the suggestions given before should apply here.

For general interval-censored failure time data, the distribution of $\{L_i, R_i\}$ or $\{U_i, V_i\}$ may depend on treatments, but the situation is not considered above. To develop a test statistic for such cases, one could apply the idea used in Section 4.2.2 to the statistics discussed in Sections 4.4.2 or 4.4.3 to adjust for the differences in the distributions of $\{U_i, V_i\}$. As mentioned for current status data, this may involve estimation of the density function of $\{U_i, V_i\}$. Another more complicated situation that also is not discussed above is that the distribution of C_i or $\{U_i, V_i\}$ is directly related to the survival function of interest rather than through treatment indicators as considered here. This is often referred to as informative censoring, and there exists very limited research on this topic. Some discussion on informative censoring is given in Section 10.5.

5

Regression Analysis of Current Status Data

5.1 Introduction

As commented before, current status data occur in many fields including animal carcinogenicity experiments, demographical studies, econometrics, epidemiological studies, and reliability studies. In some situations such as carcinogenicity experiments on occult tumors, current status data are the only information available about underlying survival variables of interest such as tumor onset time (Dinse and Lagakos, 1983). That is, the survival variables cannot be directly measured. In some other situations such as those arising from cross-sectional studies on some milestone event, current status data provide easier and more reliable information about the time to the event than complete data that give exact times to the event. An example of such situations is epidemiological studies where the event of interest is onset of certain chronic disease (Keiding, 1991; Keiding et al., 1996; Shiboski and Jewell, 1992). Another example is given by demographical studies where the event of interest can be, for instance, first pregnancy or marriage (Diamond and McDonald, 1991; Diamond et al., 1986).

There exists extensive literature about current status data in the context of animal carcinogenicity experiments and demographical studies. A detailed study of the data from these fields and discussion of the related literature are beyond the scope of this book. The objective of this chapter is regression analysis of case I or current status failure time data under the commonly used semiparametric models described in Section 1.4 with the focus on inference about regression parameters. For this, as in many other situations, the most commonly used approach is semiparametric maximum likelihood estimation. This likelihood approach is straightforward, but not easy because the likelihood is a function of finite-dimensional regression parameters and an infinite-dimensional nuisance parameter, the cumulative baseline hazard or baseline survival function. As a consequence, one has to estimate the regression parameters and nuisance parameter simultaneously. This differs from regression analysis of right-censored failure time data using the PH model

(1.4), where the partial likelihood approach can be applied. In the latter approach, for inference about the regression parameters, a partial likelihood can be derived that does not involve the nuisance parameter and whose properties can be conveniently and easily derived by martingale theory. Unfortunately, for current status data, the partial likelihood approach is not available, and one has to work with the full likelihood.

To avoid dealing with the likelihood involving an infinite-dimensional nuisance parameter, the sieve maximum likelihood method is often used in practice. The key idea behind it is to approximate the infinite-dimensional nuisance parameter by a sequence of finite-dimensional parameters, that is, the original parameter space is approximated by a sequence of increasing finite-dimensional subspaces (sieves). For the problem considered here, suppose that a semiparametric regression model is defined by regression parameter $\boldsymbol{\beta}$ and the cumulative baseline hazard function $\Lambda_0(t)$. Then the original parameter space related to $\Lambda_0(t)$ can be the collection of all nondecreasing functions, and the sieves can be, for instance, collections of nondecreasing and continuous piecewise linear functions. For any given finite sample, estimation of $\boldsymbol{\beta}$ and $\Lambda_0(t)$ can be carried out by maximizing the likelihood function over the product of the parameter spaces for $\boldsymbol{\beta}$ and the sieve. In other words, one only needs to work with a finite-dimensional parameter space with the sieve method. Another advantage of the sieve method is that the resulting estimator of the cumulative baseline function $\Lambda_0(t)$ can have a faster convergence rate than the estimator given by maximizing the full likelihood over the original parameter space (Huang and Rossini, 1997).

Another inference approach that could also avoid dealing with an infinite-dimensional nuisance parameter is to base estimation of $\boldsymbol{\beta}$ on some estimating equations. Here it is assumed that some estimating equations about $\boldsymbol{\beta}$ exist that have good properties such as unbiasedness but do not involve the nuisance parameter. A major advantage of this method is that it usually can be very easily implemented compared with the full or sieve likelihood approach. Also, it is often the case that the properties of the resulting estimates of regression parameters from this approach can be relatively easily established.

In the following sections, the inference approaches described above are discussed in details under specific semiparametric models for $\boldsymbol{\beta}$ and $\Lambda_0(t)$ or $S_0(t)$, the baseline survival function. Section 5.2 considers the PH model (1.4) and discusses the application of the semiparametric maximum likelihood approach to it. In Section 5.3, we discuss fitting the proportional odds model (1.5) to current status data. For inference, the sieve maximum likelihood approach is used along with brief description of some other approaches. The application of the additive hazards model (1.7) to the analysis of current status data is the topic of Section 5.4 with the focus on the estimating equation-based inference approach about regression parameters. Section 5.5 deals with regression analysis of discrete current status data using the grouped PH model. In Section 5.6, bibliographic notes about regression analysis of current status failure time data are provided along with some general remarks about the topic.

To define the notation for this chapter, suppose that there is a survival study that consists of n independent subjects. For the ith subject, suppose that there exist two random variables: one is the survival time of interest denoted by T_i and the other is C_i denoting the observation time on the subject, $i = 1, ..., n$. Also for subject i, suppose that there exists a vector of covariates $\boldsymbol{Z}_i$. It is assumed that the distribution of the T_i's is determined by regression parameter $\boldsymbol{\beta}$ and the baseline cumulative hazard function $\Lambda_0(t)$ or the baseline survival function $S_0(t) = \exp(-\Lambda_0(t))$. Also it is assumed that for inference about $\boldsymbol{\beta}$ and $\Lambda_0(t)$ or $S_0(t)$, only current status data are available and given in the form

$$\{\,(C_i, \delta_i = I(T_i \le C_i), \boldsymbol{Z}_i)\,;\, i \,=\, 1, ..., n\,\}\,.$$

That is, each subject is observed only once at C_i and at C_i, one knows only if the survival event of interest has occurred before or at C_i. In the following, we assume that given $\boldsymbol{Z}_i$, T_i and C_i are independent.

5.2 Analysis with the Proportional Hazards Model

This section discusses regression analysis of current status data using the PH model (1.4), the most commonly used regression model in failure time data analysis. In terms of the cumulative hazard function, the model specifies

$$\Lambda(t;\, \boldsymbol{Z}_i) \,=\, \Lambda_0(t)\, \exp(\boldsymbol{Z}_i'\,\boldsymbol{\beta})$$

for given $\boldsymbol{Z}_i$ and the likelihood function is proportional to

$$L(\boldsymbol{\beta}, \Lambda_0) \,=\, \prod_{i=1}^{n} \exp\left[-(1-\delta_i)\, e^{\boldsymbol{Z}_i'\boldsymbol{\beta}}\, \Lambda_0(C_i)\right] \left[1 - \exp(-e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \Lambda_0(C_i))\right]^{\delta_i}\,. \tag{5.1}$$

In terms of $\boldsymbol{\beta}$ and S_0, the baseline survival function, the likelihood function above has the form

$$L(\boldsymbol{\beta}, S_0) \,=\, \prod_{i=1}^{n} [S_0(C_i)]^{(1-\delta_i)\exp(\boldsymbol{Z}_i'\boldsymbol{\beta})} \left\{1 - [S_0(C_i)]^{\exp(\boldsymbol{Z}_i'\boldsymbol{\beta})}\right\}^{\delta_i}\,. \tag{5.2}$$

In the following, we first consider maximum likelihood estimation of $\boldsymbol{\beta}$ and $S_0(t)$. Two examples are then presented and followed by some discussion about asymptotic properties of the maximum likelihood estimators.

5.2.1 Maximum Likelihood Estimation

For estimation of $\boldsymbol{\beta}$ and S_0, the maximum likelihood approach maximizes the likelihood function $L(\boldsymbol{\beta}, S_0)$ given in (5.2). For this and a given set of current status data, as in the one-sample situation discussed in Sections 3.2 to 3.4, only the values of $S_0(t)$ at the observation times C_i's affect the likelihood function. Thus without loss of generality, one can focus only on the maximization of $L(\boldsymbol{\beta}, S_0)$ over all nonincreasing step functions with jumps only at the C_i's for $S_0(t)$.

Let $0 < s_1 < ... < s_m$ denote the ordered distinct time points of $\{ C_i \}_{i=1}^n$ and Ω_S the set of all baseline survival functions $S_0(t)$ that have the form

$$S_0(t) = \prod_{j \,:\, s_j \,\le\, t} e^{-\exp(\alpha_j)}\,, \tag{5.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$ are unknown parameters. Then as discussed above, to determine the maximum likelihood estimators of $\boldsymbol{\beta}$ and S_0, we only need to consider maximizing $L(\boldsymbol{\beta}, S_0)$ over $\boldsymbol{\beta}$ and S_0 in Ω_S. In this case, the log likelihood function can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n \left\{ \delta_i \log \left[1 - \prod_{j_i} e^{-\exp(\alpha_j + \boldsymbol{Z}_i'\boldsymbol{\beta})} \right] - \sum_{j_i} (1 - \delta_i) e^{\alpha_j + \boldsymbol{Z}_i'\boldsymbol{\beta}} \right\}$$

in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, where $\prod_{j_i}$ and $\sum_{j_i}$ denote the product and summation over $\{ j\,;\, s_j \le C_i \}$, respectively.

Define D_j to be the set of indices of subjects for whom $C_i = s_j$ and $\delta_i = 1$ and R_j the set of indices of subjects for whom $C_i = s_j$, $j = 1, ...m$. Let $a_j = \sum_{k=1}^j \exp(\alpha_k)$, $j = 1, ..., m$. Then the log likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be rewritten as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{j=1}^m \left\{ \sum_{i \in D_j} \log \left[\frac{1 - e^{-a_j \exp(\boldsymbol{Z}_i'\boldsymbol{\beta})}}{e^{-a_j \exp(\boldsymbol{Z}_i'\boldsymbol{\beta})}} \right] - a_j \sum_{i \in R_j} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \right\}. \tag{5.4}$$

To maximize $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$, a natural approach is to use the Newton-Raphson algorithm and for this, we need the first and second derivatives of $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$. They are

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{j=1}^m a_j \left\{ \sum_{i \in D_j} \boldsymbol{Z}_i e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \left[q(a_j, \boldsymbol{Z}_i) + 1 \right] - \sum_{i \in R_j} \boldsymbol{Z}_i e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \right\},$$

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} = e^{\alpha_j} \sum_{k=j}^m \left\{ \sum_{i \in D_k} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \left[q(a_k, \boldsymbol{Z}_i) + 1 \right] - \sum_{i \in R_k} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \right\},$$

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{j=1}^m a_j \left\{ \sum_{i \in D_j} \boldsymbol{Z}_i \boldsymbol{Z}_i' e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \left[q(a_j, \boldsymbol{Z}_i) + 1 \right.\right.$$

$$-a_j\, e^{\boldsymbol{Z}'_i\boldsymbol{\beta}} q(a_j, \boldsymbol{Z}_i)[q(a_j, \boldsymbol{Z}_i) + 1]\Big] - \sum_{i \in R_j} \boldsymbol{Z}_i \boldsymbol{Z}'_i e^{\boldsymbol{Z}'_i\boldsymbol{\beta}} \Bigg\} ,$$

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \boldsymbol{\beta}} = e^{\alpha_j} \sum_{k=j}^{m} \Bigg\{ \sum_{i \in D_k} \boldsymbol{Z}_i e^{\boldsymbol{Z}'_i\boldsymbol{\beta}} \Big[q(a_k, \boldsymbol{Z}_i) + 1$$

$$-a_j\, e^{\boldsymbol{Z}'_i\boldsymbol{\beta}} q(a_k, \boldsymbol{Z}_i)[q(a_k, \boldsymbol{Z}_i) + 1]\Big] - \sum_{i \in R_k} \boldsymbol{Z}_i e^{\boldsymbol{Z}'_i\boldsymbol{\beta}} \Bigg\} ,$$

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j^2} = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} - e^{2\alpha_j} \sum_{k=j}^{m} \Bigg\{ \sum_{i \in D_k} e^{2\boldsymbol{Z}'_i\boldsymbol{\beta}} \left[q(a_k, \boldsymbol{Z}_i) + 1 \right] q(a_k, \boldsymbol{Z}_i) \Bigg\} ,$$

and

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} = -\, e^{\alpha_j + \alpha_k} \sum_{l=k}^{m} \sum_{i \in D_l} e^{2\boldsymbol{Z}'_i\boldsymbol{\beta}} q(a_l, \boldsymbol{Z}_i) \left[q(a_l, \boldsymbol{Z}_i) + 1 \right] ,\ k > j$$

where

$$q(a_j, \boldsymbol{Z}_i) = \frac{e^{-a_j \exp(\boldsymbol{Z}'_i\boldsymbol{\beta})}}{1 - e^{-a_j \exp(\boldsymbol{Z}'_i\boldsymbol{\beta})}} ,$$

$j = 1, ..., m,\ i = 1, ..., n.$

To implement the Newton-Raphson algorithm, one needs to choose some initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ as well as a convergence criterion and to compute the inverse of a $(p+m) \times (p+m)$ matrix. One natural set of initial estimates is that given by the single point imputation approach discussed in Section 2.4.1. In terms of the convergence criterion, those discussed in Section 3.4.3 can be applied. For the inverse, a simplification can be obtained by using the fact that for a symmetric 2×2 partitioned matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} ,$$

one has

$$A^{-1} = \begin{pmatrix} A_{11|2}^{-1} & -A_{11|2}^{-1} A_{12} A_{22}^{-1} \\ -A_{22}^{-1} A_{21} A_{11|2}^{-1} & A_{22}^{-1} + A_{22}^{-1} A_{21} A_{11|2}^{-1} A_{12} A_{22}^{-1} \end{pmatrix}$$

assuming that all needed inverses exist (Rao, 1973, pp. 33), where $A_{11|2} = A_{11} - A_{12} A_{22}^{-1} A_{21}$.

Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\alpha}}_n$ denote the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ defined above. Also let $\hat{S}_n(t)$ denote the estimate of the baseline survival function given in (5.3) with $\boldsymbol{\alpha}$ replaced by $\hat{\boldsymbol{\alpha}}$ and let $\hat{\Lambda}_n(t) = -\log[\hat{S}_n(t)]$, an estimate of the baseline cumulative hazard function. To make inferences about $\boldsymbol{\beta}$, one needs to know the asymptotic distribution and an estimate of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_n$. The asymptotic normality of $\hat{\boldsymbol{\beta}}_n$ is discussed in Section 5.2.3.

For covariance estimation, a general approach is to treat $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ as a parametric likelihood function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Then one can use the observed Fisher information matrix or the submatrix of the inverse of minus the second derivative matrix of $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ corresponding to $\boldsymbol{\beta}$ to estimate the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. For the case where Z_i is dichotomous taking values 0 or 1, one can also use the following alternative due to Huang (1996).

Let $g_0(c)$ and $g_1(c)$ denote the density functions of the C_i's for subjects with $Z_i = 0$ or 1 and $\hat{g}_0(c)$ and $\hat{g}_1(c)$ their smooth kernel estimators, respectively. For given covariate Z, define

$$\hat{R}(c; Z) = \frac{\exp[-e^{Z\hat{\beta}_n}\, \hat{\Lambda}_n(c)]}{1 - \exp[-e^{Z\hat{\beta}_n}\, \hat{\Lambda}_n(c)]}\, \hat{\Lambda}_n^2(c)\, e^{2Z\hat{\beta}_n}$$

and

$$\hat{\mu}(c) = \frac{\hat{R}(c; Z = 1)\, \hat{g}_1(c)\, n_1}{\hat{R}(c; Z = 1)\hat{g}_1(c)\, n_1 + \hat{R}(c; Z = 0)\hat{g}_0(c)(n - n_1)},$$

where $n_1 = \sum_{i=1}^n Z_i$, the number of subjects with $Z_i = 1$. Then a consistent estimate of the variance of $\hat{\beta}_n$ is given by $(n\, \hat{\sigma}_n^2)^{-1}$ with

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{R}(C_i; Z_i)\, [Z_i - \hat{\mu}(C_i)]^2 \right\}. \tag{5.5}$$

The variance estimate given above is actually a consistent estimate of the information lower bound for β (see Section 5.2.3). In other words, the maximum likelihood estimator $\hat{\beta}_n$ is asymptotically efficient. A drawback of estimate (5.5) is that one has to obtain kernel estimates of $g_0(c)$ and $g_1(c)$, which requires selecting a proper bandwidth and kernel function as in Section 3.5.1. A simplification arises if $g_0(c) = g_1(c)$, which implies that C_i is independent of Z_i. In this case, it can be seen that $\hat{\mu}(c)$ and thus $\hat{\sigma}_n$ do not involve the common estimates of $g_0(c)$ and $g_1(c)$. In fact, for general $\boldsymbol{Z}_i$, if C_i is independent of $\boldsymbol{Z}_i$, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_n$ can be simply estimated by $(n\, \hat{\Sigma}_n)^{-1}$ with

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{R}(C_i; \boldsymbol{Z}_i) \left[\boldsymbol{Z}_i - \frac{\sum_{j=1}^n \boldsymbol{Z}_j \hat{R}(C_i; \boldsymbol{Z}_j)}{\sum_{j=1}^n \hat{R}(C_i; \boldsymbol{Z}_j)} \right]^{\otimes 2} \right\} \tag{5.6}$$

(Huang, 1996).

We remark that although the Newton-Raphson algorithm generally works well, the computation required could become intensive, and one may face unstable estimation problems for some data sets such as those that have a large number of different observation time points. As an alternative, for a given data set, one could maximize the likelihood function $L(\boldsymbol{\beta}, \Lambda_0)$ given in (5.1) over $\boldsymbol{\beta}$ and Λ_0 instead of $L(\boldsymbol{\beta}, S_0)$. In this case, the resulting log likelihood function has the form

$$l(\boldsymbol{\beta}, \Lambda_0) = \sum_{i=1}^{n} \left\{ \delta_i \log \left[1 - \exp(-e^{\boldsymbol{Z}_i'\boldsymbol{\beta}} \Lambda_0(C_i)) \right] - (1 - \delta_i) \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}) \Lambda_0(C_i) \right\} .$$

It can be shown that $l(\boldsymbol{\beta}, \Lambda_0)$ is concave with respect to Λ_0 for given $\boldsymbol{\beta}$, which could be used to develop some maximization procedures. Huang (1996) and Huang and Wellner (1997) studied this and suggested a two-step convex minorant algorithm. However, the convex minorant algorithm could have similar numerical problems too.

5.2.2 Two Examples

To illustrate the maximum likelihood approach presented in the previous subsection, we apply it to two examples. The first example concerns the lung tumor data described in Section 1.2.1 and analyzed in Section 4.5.1. In the second example, we consider a set of current status data arising from a study of calcification of the hydrogel intraocular lenses.

For the lung tumor data given in Table 1.3, as in Section 4.5.1, define $Z_i = 0$ for the animals in the conventional environment (CE) and 1 for those in the germ-free environment (GE). Also define the T_i's to be the occurrence times of lung tumors for the animals in the study and assume that they can be described by the PH model (1.4). For estimation of the effect of the environmental factor on tumor growth, the maximum likelihood approach gives $\hat{\beta} = 0.6934$ with estimated standard deviation equal to 0.320 based on the observed Fisher information approach. This gives a p-value of 0.03 for comparison of the two groups and a conclusion similar to that obtained in Section 4.5.1. As before, the results here suggest that the animals in the germ-free environment had significantly higher lung tumor incidence than those in the conventional environment.

Figure 5.1 presents the maximum likelihood estimates of the survival functions of times to lung tumor for animals in the two environmental groups. For comparison, the separate NPMLEs of the two survival functions given in Figure 3.1 are also included in the figure. It can be seen from Figure 5.1 that the separate estimates and the estimates given under model (1.4) seem to be reasonably close to each other, suggesting that the model (1.4) provides an acceptable approximation to the problem. The figure also suggests that the difference between the lung tumor incidences mainly occurs in later stages of the experiment.

Now we consider the data presented in Table 5.1 about calcification of the hydrogel intraocular lenses (IOL), an infrequently reported complication of cataract treatment. The study consists of 379 patients who had IOL implantation and were examined by an experienced ophthalmologist for the status of calcification. For each patient, the data give the examination time, which ranges from 0 to 37 months since the IOL implantation, and the degree of severity of IOL calcification indicated by 0 and 1. Here 0 ($\delta_i = 0$) means that no or little calcification had occurred by the time of examination, whereas 1

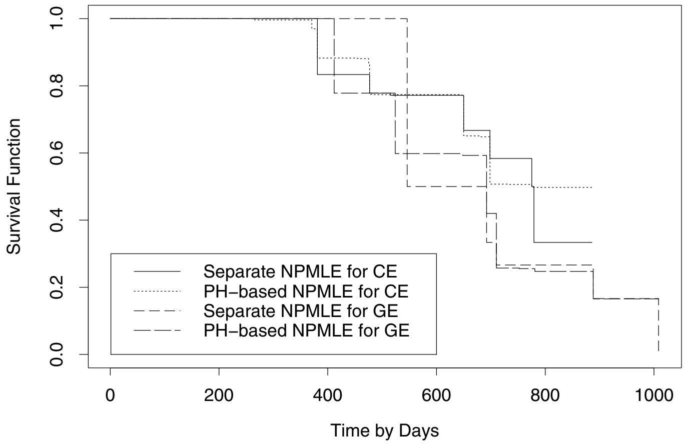

Fig. 5.1. Estimates of survival functions of time to lung tumor onset.

Table 5.1. Observed numbers of patients with ($\delta_i = 1$) and without ($\delta_i = 0$) IOL calcification at their exam times in month

Exam time	$\delta_i = 0$	$\delta_i = 1$	Exam time	$\delta_i = 0$	$\delta_i = 1$	Exam time	$\delta_i = 0$	$\delta_i = 1$
			Male patients					
1	3		2	4	1	3	5	
4	11		5	5	1	6	4	
7	9		8	7	1	9	6	1
10	6	1	11	15	1	12	5	1
13	7	1	14	8	2	15	5	
16	5	3	17	6	1	18	1	1
19	4		22	1		24	3	
26	3		28	2		30	1	
32	1							
			Female patients					
2	9		3	6	1	4	7	1
5	10		6	10		7	13	1
8	12	1	9	12	2	10	15	3
11	16	5	12	19	1	13	17	3
14	7		15	10	5	16	14	3
17	7	2	18	2		19	3	1
20	2	1	21	1		22	1	
23	1		24	2		25	1	
26	2		27		1	29	2	
30	1	1	31	1		33		1
37	1							

($\delta_i = 1$) means that there existed mild or serious calcification, or calcification had already occurred at the time of examination. Also given in the table is the gender of each patient. We note that the study consists of 142 males and 237 females. The original study, analyzed by Yu et al. (2001) and Xue et al. (2004), includes more detailed classifications about the severity of IOL calcification and some other covariates. The objective here is to estimate the gender effect on the IOL calcification and to evaluate if the risks or hazards of IOL calcification between male and female patients are identical.

Define T_i to be the time to the occurrence of IOL calcification for patient i, $i = 1, ..., 379$, and suppose that the T_i's follow the PH model (1.4). Then for the T_i's, we only have current status data available. Define $Z_i = 0$ for female patients and 1 otherwise, $i = 1, ..., 379$. The maximum likelihood approach gives $\hat{\beta} = -0.2241$ and its estimated standard error is 0.295 based on the observed Fisher information approach. Based on the standard normal distribution, this yields a p-value of 0.448 for testing $\beta = 0$ and suggests that there is no significant difference between male and female patients in terms of the time to IOL calcification. To give a graphical comparison, Figure 5.2 displays the separate NPMLEs of survival functions of the time to IOL calcification for male and female patients obtained using the algorithm given in Section 3.2. Also included in the figure are the maximum likelihood estimates of the same survival functions derived under the PH model. Figure 5.2 confirms the conclusion obtained above and also indicates that the PH model seems reasonable.

For the data set considered here, it can be easily verified that the observation or examination times C_i's seem to be independent of the gender factor.

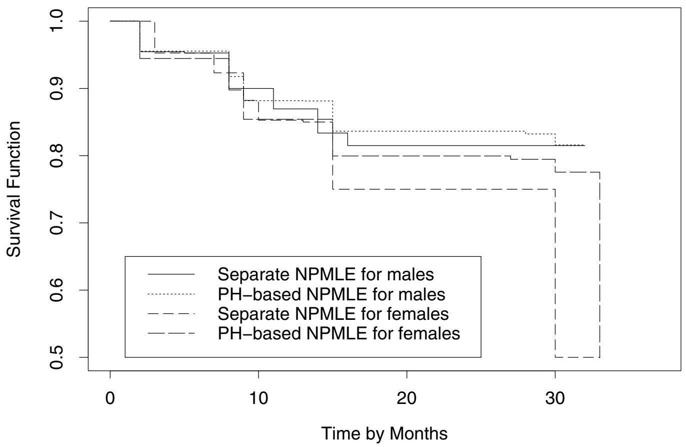

Fig. 5.2. Estimates of survival functions of time to IOL calcification.

More discussion on this is given in Section 5.4.2. This suggests that one could also use formula (5.6) for variance estimation. It gives an estimated standard error of 0.3117, similar to that given above by the observed Fisher information approach.

5.2.3 Asymptotics

This subsection discusses some asymptotic properties of the maximum likelihood estimators derived in Section 5.2.1. Suppose that $S_0(t)$ is continuous and the $\boldsymbol{Z}_i$'s are bounded. Then under some regularity conditions, Huang (1996) shows that both $\hat{\boldsymbol{\beta}}_n$ and $\hat{S}_n(t)$ are consistent. In particular, if the distribution function $G(c)$ of the C_i's is discrete, then one has

$$\hat{S}_n(t) \to S_0(t)$$

almost surely at all the mass points of G as $n \to \infty$. If $G(c)$ is continuous, as $n \to \infty$, one has

$$\sup_{0 \leq t < \infty} |\, \hat{S}_n(t) - S_0(t) \,| \to 0$$

almost surely. This implies that as $n \to \infty$,

$$\sup_{0 \leq t \leq C_0} |\, \hat{\Lambda}_n(t) - \Lambda_0(t) \,| \to 0$$

almost surely for any finite constant C_0.

As discussed in Section 3.6, for case I or II interval-censored failure time data, the convergence rate of maximum likelihood estimators can be slower than the usual $\sqrt{n}$-convergence rate. This is also true for $\hat{\Lambda}_n$ under the situation considered here. Specifically, suppose that the joint distribution $G(c, \boldsymbol{z})$ of the C_i's and $\boldsymbol{Z}_i$'s has bounded second order partial derivative with respect to c. Then under some regularity conditions, we have

$$\left\{ \int_{\tau_0}^{\tau_1} \left[\hat{\Lambda}_n(t) - \Lambda_0(t) \right]^2 d\, G_n(t) \right\}^{1/2} = O_p(n^{-1/3}) \,,$$

where $G_n(c)$ denotes the marginal empirical distribution of the C_i's. That is, $\hat{\Lambda}_n$ has only $n^{1/3}$-convergence rate, which is the same as that of the NPMLE of a distribution function when only interval-censored data are available. As seen below, however, $\hat{\boldsymbol{\beta}}_n$ still has the usual $\sqrt{n}$-convergence rate.

For the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$, in addition to the assumptions specified above, we also need that $G(c)$ has bounded support $I_C = [\tau_0, \tau_1]$ with $\tau_0 > 0$. Furthermore, $S_0(t)$ has strictly positive and continuous density on I_C. Then Huang (1996) proves that under some regularity conditions, as $n \to \infty$,

$$\sqrt{n} \, (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to N(0, \Sigma^{-1})$$

in distribution. In the above, Σ is the information matrix for $\boldsymbol{\beta}$ and can be estimated by $\hat{\sigma}_n^2$ or $\hat{\Sigma}_n$ given in (5.5) or (5.6) for finite samples if Z_i is dichotomous or C_i is independent of $\boldsymbol{Z}_i$, respectively. The result here says that $\hat{\boldsymbol{\beta}}_n$ not only has the usual $\sqrt{n}$-convergence rate, but also is asymptotically efficient as its asymptotic variance achieves the information lower bound.

5.3 Analysis with the Proportional Odds Model

In this section, we discuss use of the proportional odds model (1.5) for regression analysis of current status data. Define $H(t) = -\operatorname{logit}[S_0(t)]$ as before. Then the conditional survival function given covariates $\boldsymbol{Z}$ has the form

$$S(t; \boldsymbol{Z}) = \frac{1}{1 + \exp[H(t) + \boldsymbol{Z}'\boldsymbol{\beta}]}$$

and the likelihood function is proportional to

$$L(\boldsymbol{\beta}, H) = \prod_{i=1}^{n} \left\{ \frac{\exp[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]}{1 + \exp[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} \right\}^{\delta_i} \left\{ \frac{1}{1 + \exp[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} \right\}^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left\{ \frac{\exp\{\delta_i\,[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]\}}{1 + \exp[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} \right\} .$$

This gives the log likelihood as

$$l(\boldsymbol{\beta}, H) = \sum_{i=1}^{n} \{\delta_i\,\left[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}\right] - \log\{1 + \exp[H(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]\}\} .$$

5.3.1 Sieve Maximum Likelihood Estimation

For estimation of $\boldsymbol{\beta}$ along with H, we consider the sieve maximum likelihood estimation approach. As discussed before, the key idea behind this approach is to approximate $H(t)$, the infinite-dimensional nuisance parameter, by a sequence of finite-dimensional parameters or functions known up to finite-dimensional parameters. Let $H_{\boldsymbol{\theta}}(t)$ denote a function that is known up to a k-dimensional parameter $\boldsymbol{\theta}$ and can be used to approximate $H(t)$. The selection of $H_{\boldsymbol{\theta}}(t)$ is discussed below. Then estimation of $\boldsymbol{\beta}$ and H involves maximization of the approximate parametric log likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\theta}) = l(\boldsymbol{\beta}, H = H_{\boldsymbol{\theta}})$ over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ denote the values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ that maximize $l(\boldsymbol{\beta}, \boldsymbol{\theta})$. To determine these estimators, one needs the first partial derivatives of the log likelihood function. For $i = 1, ..., n$, define

$$E_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\exp[H_{\boldsymbol{\theta}}(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]}{1 + \exp[H_{\boldsymbol{\theta}}(C_i) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} .$$

Then the first partial derivatives are

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{Z}_i \, [\delta_i - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})]$$

and

$$U_{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{H}_{\boldsymbol{\theta}}^{(1)}(C_i) \, [\delta_i - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})] \, ,$$

where $\boldsymbol{H}_{\boldsymbol{\theta}}^{(1)}(t) = \partial H_{\boldsymbol{\theta}}(t)/\partial \boldsymbol{\theta}$, the k-dimensional vector of derivatives of $H_{\boldsymbol{\theta}}(t)$. Hence the approximate maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be obtained by solving the score equations

$$U(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ U_{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) \end{pmatrix} = 0 \, .$$

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be estimated by the observed Fisher information matrix. For this, one needs the second partial derivatives and we have

$$I_{11} = -\frac{\partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i' E_i(\boldsymbol{\beta}, \boldsymbol{\theta})[1 - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})] \, ,$$

$$I_{12} = -\frac{\partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{H}_{\boldsymbol{\theta}}^{(1)'}(C_i) E_i(\boldsymbol{\beta}, \boldsymbol{\theta})[1 - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})] \, ,$$

and

$$I_{22} = -\frac{\partial U_{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \Big\{ \boldsymbol{H}_{\boldsymbol{\theta}}^{(1)}(C_i) \boldsymbol{H}_{\boldsymbol{\theta}}^{(1)'}(C_i) E_i(\boldsymbol{\beta}, \boldsymbol{\theta})[1 - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})]$$

$$- \, [\delta_i - E_i(\boldsymbol{\beta}, \boldsymbol{\theta})] H_{\boldsymbol{\theta}}^{(2)}(C_i) \Big\} \, ,$$

where $\boldsymbol{H}_{\boldsymbol{\theta}}^{(2)}(t) = \partial \boldsymbol{H}_{\boldsymbol{\theta}}^{(1)}(t)/\partial \boldsymbol{\theta}$, a $k \times k$ matrix. Thus the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated by the submatrix of the inverse of

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{12}' & I_{22} \end{bmatrix}$$

corresponding to $\boldsymbol{\beta}$.

To apply the sieve estimation approach described above, one needs to choose $H_{\boldsymbol{\theta}}(t)$ for a given sample and many choices exist. It is apparent that the simplest one is a piecewise constant function and another commonly used choice is a spline function. Here we focus on the piecewise constant functions. Specifically, suppose that the distribution function of the C_i's has bounded support $[0, \tau]$ and that $0 = t_0 < t_1 < ... < t_k = \tau$ is a partition of the interval $[0, \tau]$. Then a step function over $[0, \tau]$ can be expressed as

$$H_{\boldsymbol{\theta}}(t) = \sum_{j=1}^{k} \theta_j \, I_j(t) \,, \tag{5.7}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$ and $I_j(t) = I(t_{j-1} < t \leq t_k)$, the indicator function for the jth interval $(t_{j-1}, t_j]$. Given the t_j's, $H_{\boldsymbol{\theta}}$ is completely determined by parameter $\boldsymbol{\theta}$. Using the function given in (5.7), we have

$$\boldsymbol{H}^{(1)}_{\boldsymbol{\theta}}(t) = \begin{pmatrix} I_1(t) \\ \cdot \\ \cdot \\ \cdot \\ I_k(t) \end{pmatrix}$$

and $\boldsymbol{H}^{(2)}_{\boldsymbol{\theta}}(t) = 0$.

In practice, the number of partition intervals or steps of $H_{\boldsymbol{\theta}}$, k, needs to increase with the sample size n because otherwise $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ will not be consistent. Rossini and Tsiatis (1996) investigated the sieve estimation approach under model (5.7) and the asymptotic properties of the estimators. They show that under some regularity conditions, $\hat{\boldsymbol{\beta}}$ and $H_{\hat{\boldsymbol{\theta}}}$ are consistent estimators of $\boldsymbol{\beta}$ and H if $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$. Furthermore, if k increases at a rate $k(n) = O(n^{\alpha})$ with $1/4 < \alpha < 1$, then $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a normal random variable with mean zero and the variance-covariance achieving the information lower bound. That is, $\hat{\boldsymbol{\beta}}$ is asymptotically efficient.

It is easy to see that for larger k, more computational effort is needed, but a better approximation is obtained. Based on the results given above, one can choose k to be the smallest integer above $n^{1/4}$, but for small n, one may want to choose larger k since the results may not be stable otherwise. The largest integer for k is apparently the number of all different observation times C_i. In this case, the sieve maximum likelihood estimation procedure is equivalent to the usual maximum likelihood estimation procedure. For given k, one simple way to choose the t_j's is to use the equally spaced partition. Another way is to choose intervals such that they contain approximately equal numbers of observation times. More comments on k and the t_j's are given in the next subsection through examples.

5.3.2 Illustrations

For comparison, we apply the sieve maximum likelihood approach discussed above to the two examples discussed in Section 5.2.2. First consider the lung tumor data and let the T_i's and Z_i's be defined as in Section 5.2.2. To apply the sieve maximum likelihood approach, we use the step function given in (5.7) to approximate $H(t)$ and divide the range of animal death times, $(0, 1000)$, equally for the selection of the t_j's. For the number of partition points, several values of k are used starting with $k = 4$, the smallest integer larger than $144^{1/4}$. Table 5.2 gives estimated regression parameters and their estimated standard errors. It can be seen from the table that both point and variance estimation of the regression parameter is quite stable and the resulting p-values for testing $\beta = 0$ are around 0.001. This indicates that the animals in the germ-free environment had significantly higher rate of occurrence of lung tumor than those in the conventional environment, and the conclusion is similar to that obtained using the PH model. The smaller p-value here suggests that the difference between the tumor occurrence rates of the animals in the two environments is more significant in terms of the logit difference of the survival functions than the log-log difference of the survival functions or the log difference of the cumulative hazard functions.

To investigate the effect of the partition or selection of the t_j's on the analysis, we repeated the analysis above by choosing intervals such that they contain roughly equal numbers of observations. For $k = 4$ (each interval contains 36 observations), for example, the approach results in $\hat{\beta} = 1.4907$ with estimated standard error equal to 0.4637. The results are similar to those given in Table 5.2.

For the application of the sieve maximum likelihood approach to the calcification data, we use the step function given in (5.7) for approximation as with the lung tumor data and partition the range of observation times, $(0, 37)$, evenly into k intervals. Using the same T_i's and Z_i's as defined in Section 5.2.2, we obtain estimates of the regression parameter and their estimated standard errors for different k. The results are given in Table 5.3 and in this example, we start with $k = 5$, the smallest integer greater than $379^{1/4}$. As with the lung tumor data, the results are quite stable. The resulting p-values for comparing the calcification occurrence rates between the male and female patients are around 0.4 and are consistent with that obtained under the PH model.

Because there exist many tied observation times in the calcification data, it is impossible to put all the observation times into adjacent groups with equal numbers in each group. For $k = 5$, choosing intervals containing approximately equal numbers of observations times, we divide $(0, 37]$ into five intervals, containing 78, 90, 62, 90 and 58 observations, respectively. With

Table 5.2. Estimates of regression parameter for lung tumor data

No. of partitions (k)	$\hat{\beta}$	SD of $\hat{\beta}$	No. of partitions (k)	$\hat{\beta}$	SD of $\hat{\beta}$
4	1.5620	0.4516	5	1.5299	0.4637
6	1.5575	0.4707	7	1.6866	0.4685
8	1.5733	0.4819	10	1.5597	0.4684

Table 5.3. Estimates of regression parameter for calcification data

No. of partitions (k)	$\hat{\beta}$	SD of $\hat{\beta}$	No. of partitions (k)	$\hat{\beta}$	SD of $\hat{\beta}$
5	-0.2360	0.3375	6	-0.2748	0.3357
7	-0.2573	0.3370	8	-0.2620	0.3388
9	-0.2386	0.3397	12	-0.2698	0.3373

the use of the same sieve function as above, the sieve maximum likelihood approach gives $\hat{\beta} = -0.2782$ with an estimated standard deviation of 0.3353, similar to the results presented in Table 5.3.

5.3.3 Discussion

Note that the baseline log-odds function $H(t)$ is nondecreasing and continuous assuming that $S_0(t)$ is continuous, whereas the piecewise constant function (5.7) is apparently not. If the focus is on regression parameters, this should not affect the inference. On the other hand, if one is interested in estimation of $H(t)$ or $S_0(t)$, one may want to put order restriction on the θ_j's in (5.7). Alternatively, instead of (5.7), one may want to use

$$H_{\boldsymbol{\theta}}(t) = \sum_{j=1}^{k} \left[\frac{\theta_j - \theta_{j-1}}{t_j - t_{j-1}} \, t - \frac{\theta_j t_{j-1} - \theta_{j-1} t_j}{t_j - t_{j-1}} \right] I_j(t) \qquad (5.8)$$

for given $t_0 < t_1 < ... < t_k$, where $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_k)'$ with $\theta_0 \leq \theta_1 ... \leq \theta_k$. It is apparent that this is a continuous nondecreasing function. Huang and Rossini (1997) studied the sieve maximum likelihood estimation approach using the function given in (5.8) for general interval-censored data. They show that the resulting estimators have similar asymptotic properties to those of the estimators discussed in Section 5.3.1. Of course, one can also use some spline functions, which are commonly used to approximate unknown smooth or continuous functions (Shen, 1998).

Instead of the sieve maximum likelihood estimation approach, one can naturally use the maximum likelihood estimation approach that directly maximizes $l(\boldsymbol{\beta}, H)$ for estimation of $\boldsymbol{\beta}$ and H (Dinse and Lagakos, 1983; Huang, 1995). One advantage of the former approach is that the resulting estimators may have a faster convergence rate than the maximum likelihood estimators under certain smoothness assumptions (Huang and Rossini, 1997). Also the sieve maximum likelihood estimators may be easier to compute than the maximum likelihood estimators because fewer parameters are involved for the former. The main disadvantage of the sieve estimation approach is that one has to choose the partition and finite-dimensional functions, or more generally the finite-dimensional parameter or function space. For a given finite sample, different partition and function space may result in different analysis results. Also for each selected function space, one needs to investigate properties of the resulting estimates of parameters because no general theory exists.

5.4 Analysis with the Additive Hazards Model

As mentioned before, the additive hazards model (1.7) is another commonly used regression model in survival analysis in addition to the PH model and the proportional odds model. For fitting model (1.7) to current status data, we focus on inference about regression parameters and consider a simple estimating equation approach first considered by Lin, Oakes, and Ying (1998).

5.4.1 Estimation of Regression Parameters

To make inference about regression parameter $\boldsymbol{\beta}$ in model (1.7), we assume that given $\boldsymbol{Z}_i$, the conditional hazard function of C_i is given by

$$\lambda_i^c(t; \boldsymbol{Z}_i) \,=\, \lambda_c(t)\, \exp(\boldsymbol{Z}_i' \gamma)\,. \tag{5.9}$$

In this model, $\lambda_c(t)$ is an unknown baseline hazard function like $\lambda_0(t)$ in model (1.7) and $\boldsymbol{\gamma}$ is the regression parameters representing the effect of covariates on C_i. That is, C_i follows the PH model.

Define the counting process $N_i(t) \,=\, I\{C_i \leq \min(T_i, t)\}$. Then N_i jumps by one at time t if and only if $C_i \,=\, t$ and $T_i \,\geq\, t$. A subject with N_i can be regarded as censored if $C_i \,=\, t$ and it is found that $T_i \,<\, t$, meaning that the subject is at risk at t if and only if $C_i \,\geq\, t$. It follows that the intensity function $dN_i(t)$ has the form

$$d\Lambda_i^*(t; \boldsymbol{Z}_i) \,=\, Y_i(t)\, \exp(-t\boldsymbol{Z}_i'\boldsymbol{\beta} \,+\, \boldsymbol{Z}_i'\gamma)\, d\Lambda_0^*(t)\,, \tag{5.10}$$

where $Y_i(t) \,=\, I(C_i \geq t)$, $\Lambda_0^*(t) \,=\, e^{-\Lambda_0(t)}\, d\Lambda_c(t)$ with $\Lambda_0(t) \,=\, \int_0^t \lambda_0(s) ds$ and $\Lambda_c(t) \,=\, \int_0^t \lambda_c(s) ds$. This suggests that the process

$$M_i(t) \,=\, N_i(t) \,-\, \int_0^t Y_i(s)\, \exp(\boldsymbol{Z}_i'\gamma \,-\, t\boldsymbol{Z}_i'\boldsymbol{\beta})\, d\Lambda_0^*(s) \tag{5.11}$$

is a martingale, $i \,=\, 1, ..., n$, and that one can apply the partial likelihood approach to model (5.10) for inference about $\boldsymbol{\beta}$.

Define

$$S_{\boldsymbol{\beta}}^{(j)}(t; \boldsymbol{\beta}, \gamma) \,=\, \sum_{i=1}^n Y_i(t)\, (t\boldsymbol{Z}_i)^{(j)}\, \exp(\boldsymbol{Z}_i'\gamma \,-\, t\boldsymbol{Z}_i'\boldsymbol{\beta})\,,$$

where $(t\boldsymbol{Z}_i)^{(0)} \,=\, 1$ and $(t\boldsymbol{Z}_i)^{(1)} \,=\, t\boldsymbol{Z}_i$, $j \,=\, 0, 1$. The application of the partial likelihood approach yields the partial score function

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \gamma) \,=\, \sum_{i=1}^n \int_0^\infty \left[t\boldsymbol{Z}_i \,-\, \frac{S_{\boldsymbol{\beta}}^{(1)}(t; \boldsymbol{\beta}, \gamma)}{S_{\boldsymbol{\beta}}^{(0)}(t; \boldsymbol{\beta}, \gamma)} \right] dN_i(t)$$

for estimation of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$. For estimation of $\boldsymbol{\gamma}$, one can obtain a partial score function similar to $U_{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{\gamma})$ under model (5.10). However, a more efficient approach is to apply the partial likelihood approach to model (5.9) because the C_i's are always available. This gives the partial score function

$$U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \int_0^{\infty} \left[\boldsymbol{Z}_i - \frac{S_{\boldsymbol{\gamma}}^{(1)}(t;\boldsymbol{\gamma})}{S_{\boldsymbol{\beta}}^{(0)}(t;\boldsymbol{\gamma})} \right] dI(C_i \leq t)\,,$$

where

$$S_{\boldsymbol{\gamma}}^{(j)}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} Y_i(t)\, \boldsymbol{Z}_i^{(j)}\, e^{\boldsymbol{Z}'_i \boldsymbol{\gamma}}$$

with $\boldsymbol{Z}_i^{(0)} = 1$ and $\boldsymbol{Z}_i^{(1)} = \boldsymbol{Z}_i$, $j = 0, 1$.

Let $\hat{\boldsymbol{\gamma}}$ denote the partial likelihood estimator of $\boldsymbol{\gamma}$ given by the solution to $U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = 0$. Then one can estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ defined as the solution to $U_{\boldsymbol{\beta}}(\boldsymbol{\beta};\hat{\boldsymbol{\gamma}}) = 0$. Both $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ can be easily obtained by the Newton-Raphson algorithm. It can be shown that under some regularity conditions, $\hat{\boldsymbol{\beta}}$ is consistent as well as $\hat{\boldsymbol{\gamma}}$ (Kalbfleisch and Prentice, 2002; Lin, Oakes, and Ying, 1998). Also for large n, the distributions of $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{n}\,(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ can be approximated by the multivariate normal distribution with mean zero and variance-covariance matrices

$$I_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}}) - I_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})\, I_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})\, I_{\gamma}^{-1}(\hat{\boldsymbol{\gamma}})\, I'_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})\, I_{\boldsymbol{\beta}}^{-1}(\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\gamma}})$$

and $I_{\gamma}^{-1}(\hat{\boldsymbol{\gamma}})$, respectively, where

$$I_{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{\gamma}) = \frac{1}{n}\, \frac{\partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{\gamma})}{\partial \boldsymbol{\beta}}\,,\quad I_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{\beta};\boldsymbol{\gamma}) = -\frac{1}{n}\, \frac{\partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta};\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\,,$$

and

$$I_{\gamma}(\boldsymbol{\gamma}) = -\frac{1}{n}\, \frac{\partial U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\,.$$

It is seen that the approach here essentially transforms the analysis problem to regression analysis of right-censored failure time data using the PH model, which can be quite easily performed. In consequence, one can apply existing software for the PH model to determine the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$.

Note that unlike the approaches discussed in Sections 5.2 and 5.3, the estimation procedure described above requires that the C_i's follow the PH model. This may be restrictive and could result in biased estimates of regression parameters if the model is incorrect. On the other hand, we note that model (5.9) can be easily verified because complete data are always available for the C_i's. For the case where Z_i takes values 0 and 1, for example, a simple approach for assessing the model (5.9) is to obtain and compare separate estimates of the survival functions for subjects with $Z_i = 0$ and 1, respectively, and the estimates given under the model (5.9).

5.4.2 Illustrations

Again, we consider the data sets analyzed in Section 5.2.2 as well as in Section 5.3.2 with the T_i's and Z_i's defined in the same way. For the lung tumor data, as discussed in Section 4.5.1, the model (5.9) seems reasonable for description of the animal death time, C_i, and the result there gave $\hat{\gamma} = -1.9627$ with estimated standard error equal to 0.243. This indicates that the animals in the germ-free environment had much significantly longer survival times than those in the conventional environment. The application of the inference procedure described in the previous subsection gives $\hat{\beta} = 0.0007$ and an estimated standard deviation 0.0004. This corresponds with a p-value of 0.085 for comparing the lung tumor rates between the animals in the two environments, which is less significant than the results given in in the previous sections. It should be noted that one cannot directly compare $\hat{\beta}$ obtained here to those given before because they represent different quantities. For example, $\hat{\beta}$ here gives the estimated difference between the two hazard functions for animals in the two environments. In contrast, the estimate given in Section 5.2.2 represents the estimated log of the ratio of the two hazard functions, while $\hat{\beta}$ in Section 5.3.2 estimates the logit difference between the two corresponding survival functions.

Now we consider the application of the inference procedure given in the previous subsection to the calcification data. Assume that the T_i's follow the additive hazards model (1.7). Using the inference procedure of Section 5.4.1, we obtain $\hat{\beta} = -0.0015$ with estimated standard error equal to 0.005. This in-

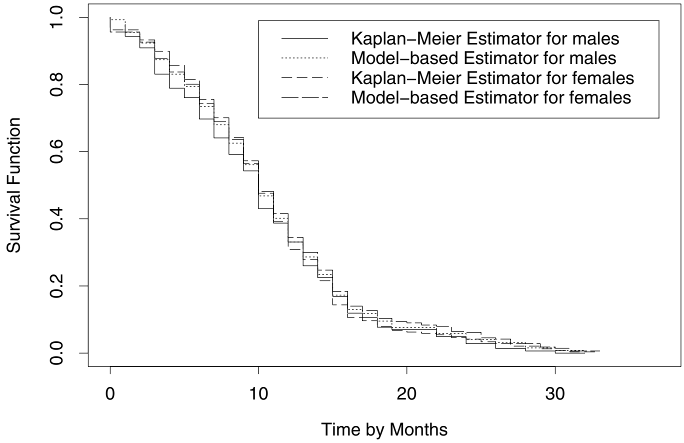

Fig. 5.3. Estimates of survival functions of observation times for male and female patients.

dicates that there is no significant difference between the hazards of developing IOL calcification for the male and female patients, and the result is consistent with those obtained using the PH and the proportional odds models. To check the appropriateness of model (5.9), as with Figure 4.1 for the lung tumor data, Figure 5.3 presents the estimates of the survival functions of examination times for male and female patients obtained under model (5.9) along with their corresponding Kaplan-Meier estimators. It suggests that model (5.9) seems reasonable and actually, the distribution of examination times seems to be independent of the gender. This is confirmed by $\hat{\gamma} = 0.0372$ with estimated standard error of 0.106.

5.4.3 Discussion

Compared with the approaches discussed in Sections 5.2 and 5.3, the estimating equation approach considered in this section has the advantage that it does not involve estimation of the baseline cumulative hazard function $\Lambda_0(t)$. This makes estimation of regression parameters much easier. Also, implementation of the approach here is much simpler because one can make use of some existing software for the PH model with right-censored failure time data. Another advantage is that the derivation of asymptotic properties of resulting estimates of regression parameters is much easier again because one essentially deals with the PH model with right-censored data rather than current status data.

In general, an estimating equation approach can have the disadvantage that it may not be as efficient as the maximum likelihood or sieve maximum likelihood approach. This is true for the method described in Section 5.4.1 because the distribution of the censoring time for $N_i(t)$ involves regression parameter $\boldsymbol{\beta}$ (Lin, Oakes, and Ying, 1998; Martinussen and Scheike, 2002b). That is, the censoring times, which are usually assumed to be independent of survival times given covariates in the application of the partial likelihood approach, are informative. In other words, the method discussed here represents a trade-off between simplicity and efficiency and the amount of efficiency loss depends on specific situations.

As an alternative to the estimating equation approach for estimation of $\boldsymbol{\beta}$, one can apply the maximum likelihood approach or the sieve maximum likelihood approach as in Section 5.2 or 5.3. For example, Ghosh (2001) investigated the maximum likelihood approach for fitting model (1.7) to current status data. Another alternative is to directly derive and use the efficient score function for $\boldsymbol{\beta}$, which has the form

$$U_E(\boldsymbol{\beta};\Lambda_0) \;=\; \sum_{i=1}^{n} \int_0^\infty \left[t\boldsymbol{Z}_i \;-\; \frac{S_E^{(1)}(t;\boldsymbol{\beta},\Lambda_0)}{S_E^{(0)}(t;\boldsymbol{\beta},\Lambda_0)} \right]$$

$$\times \left\{ \frac{\exp[-\Lambda_0(t) - t\boldsymbol{Z}_i'\boldsymbol{\beta}]}{1-\exp[-\Lambda_0(t) - t\boldsymbol{Z}_i'\boldsymbol{\beta}]}\, dN_i^*(t) \;-\; dN_i(t) \right\}$$

(Martinussen and Scheike, 2002b). In the expression above, $N_i^*(t) = I(C_i \leq t) - N_i(t)$ and

$$S_E^{(j)}(t;\boldsymbol{\beta},\Lambda_0) = \sum_{i=1}^{n} Y_i(t)\,\alpha_i^c(t;\boldsymbol{Z}_i)\,\frac{\exp[-\Lambda_0(t) - t\boldsymbol{Z}_i'\boldsymbol{\beta}]}{1-\exp[-\Lambda_0(t) - t\boldsymbol{Z}_i'\boldsymbol{\beta}]}\,(t\boldsymbol{Z}_i)^{(j)}\,,$$

where $\alpha_i^c(t;\boldsymbol{Z}_i)$ denotes the hazard function of C_i given $\boldsymbol{Z}_i$, which does not have to satisfy model (5.9). To apply this approach, one needs first to estimate both $\Lambda_0(t)$ and $\alpha_i^c(t;\boldsymbol{Z}_i)$ and then can estimate $\boldsymbol{\beta}$ by the solution to $U_E(\boldsymbol{\beta};\Lambda_0) = 0$ with both $\Lambda_0(t)$ and $\alpha_i^c(t;\boldsymbol{Z}_i)$ replaced by their estimates. Martinussen and Scheike (2002b) show that such defined estimator is consistent and asymptotically has a multivariate normal distribution with covariance matrix reaching the information lower bound.

It is straightforward to generalize both the estimating equation approach and the efficient score function approach to situations where covariates are time-dependent (Lin, Oakes, and Ying, 1998; Lin and Ying, 1997; Martinussen and Scheike, 2002b).

5.5 Analysis with the Grouped Proportional Hazards Models

This section deals with situations where the survival time of interest takes only finite discrete values $0 < s_1 < ... < s_{m+1}$ with the survival probability at s_{m+1} equal to zero. This can arise if the time axis is divided into $m+1$ time intervals, and the survival event can only be observed to occur within certain interval or intervals. It can also arise if study subjects can only be observed at the s_j's due to, for example, study design. As in Section 5.2.1, for each j $(1 \leq j \leq m)$, let D_j denote the set of indices of subjects for whom $C_i = s_j$ and $\delta_i = 1$ and R_j the set of indices of subjects for whom $C_i = s_j$. Then the likelihood function is proportional to

$$L(\boldsymbol{\beta}, S_0) = \prod_{j=1}^{m} \prod_{i \in D_j} [1 - S(s_j;\boldsymbol{Z}_i)] \prod_{i \in R_j - D_j} S(s_j;\boldsymbol{Z}_i)$$

with respect to regression parameter $\boldsymbol{\beta}$ and the baseline survival function $S_0(t)$. In the following, we assume that $S(s_j;\boldsymbol{Z}_i)$ is given by the grouped PH model (1.10) or (1.11). This gives the log likelihood function $l(\boldsymbol{\beta},\boldsymbol{\alpha}) = \log\, L(\boldsymbol{\beta}, S_0)$ as

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{j=1}^{m} \left\{ \sum_{i \in D_j} \log\left[\frac{1-\prod_{k=1}^{j} e^{-\exp(\alpha_k + \boldsymbol{Z}_i'\boldsymbol{\beta})}}{\prod_{k=1}^{j} e^{-\exp(\alpha_k + \boldsymbol{Z}_i'\boldsymbol{\beta})}}\right] - \sum_{i \in R_j} \sum_{k=1}^{j} e^{\alpha_k + \boldsymbol{Z}_i'\boldsymbol{\beta}} \right\}$$

in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$ defined in model (1.11).

5.5.1 Maximum Likelihood Estimation of Parameters

Define $a_j = \sum_{k=1}^{j} \exp(\alpha_k)$, $j = 1, ..., m$. The log likelihood function above can then be rewritten as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{j=1}^{m} \left\{ \sum_{i \in D_j} \log \left[\frac{1 - e^{-a_j \exp(\boldsymbol{Z}'_i \boldsymbol{\beta})}}{e^{-a_j \exp(\boldsymbol{Z}'_i \boldsymbol{\beta})}} \right] - a_j \sum_{i \in R_j} e^{\boldsymbol{Z}'_i \boldsymbol{\beta}} \right\} ,$$

which is the same as the log likelihood function given in (5.4).

For the maximum likelihood estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, as in Section 5.2.1, one can apply the Newton-Raphson algorithm using the first and second derivatives of $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ given there. Although the estimation procedure is the same for the situations discussed here and in Section 5.2, it is apparent that the problems considered are different. In particular, the derivation of asymptotic properties of the maximum likelihood estimators obtained here is straightforward and follows the standard likelihood theory for parametric models.

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ denote the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ given by the solution to

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = 0 \text{ and } \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} = 0 .$$

As discussed above, they can be obtained by the Newton-Raphson algorithm. Their variance-covariance matrix can be estimated by the inverse of the observed Fisher information matrix

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$, where

$$I_{11} = -\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} , \; I_{12} = I'_{21} = -\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \boldsymbol{\beta}} \right) , \; I_{22} = -\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} \right) .$$

In many situations, tests about $\boldsymbol{\beta}$ are of particular interest, and for this purpose, the score test procedure can be applied. For example, consider the two-sample survival comparison where covariate Z_i is defined as a dichotomous variable taking values 0 and 1. In this case, the survival comparison is equivalent to testing $\beta = 0$ and the score function for β has the form

$$U_\beta(\beta, \boldsymbol{\alpha}) = \frac{\partial l(\beta, \boldsymbol{\alpha})}{\partial \beta} = \sum_{j=1}^{m} a_j \, e^{\beta} \left(\frac{d_{j1}}{1 - e^{-a_j \exp(\beta)}} - n_{j1} \right) ,$$

where $d_{j1} = \sum_{i \in D_j} Z_i$ and $n_{j1} = \sum_{i \in R_j} Z_i$. For large n, the distribution of this score statistic can be approximated by the normal distribution with mean zero and variance

$$\sigma^2(\beta, \boldsymbol{\alpha}) = I_{11} - I_{12} I_{22}^{-1} I_{21} .$$

Thus the test of $\beta = 0$ can be based on the statistic

$$\frac{U_\beta^2(0, \hat{\boldsymbol{\alpha}}_0)}{\sigma^2(0, \hat{\boldsymbol{\alpha}}_0)} ,$$

which is approximately $\chi^2_{(1)}$, where $\hat{\boldsymbol{\alpha}}_0$ denotes the maximum likelihood estimator of $\boldsymbol{\alpha}$ at $\beta = 0$.

5.5.2 Two Examples

To illustrate the inference procedure described above and compare it with the other procedures, we consider the grouped lung tumor data obtained by dividing the observation time period in the lung tumor data discussed in the previous sections into 10 equally spaced intervals. That is, it is assumed that each animal was only observed to die in one of the intervals $(0, 100]$, $(100, 200]$, ... , $(900, 1000]$ with or without lung tumor. Then we have $m = 10$ and s_j can be taken to be any number within the interval $(100(j-1), 100j]$ with s_{11} being any number larger than 1000, $j = 1, ..., 10$. As in the previous sections, define $Z_i = 0$ for the animals in the conventional environment and 1 for the animals in the germ-free environment.

The application of the inference procedure gives $\hat{\beta} = 0.8424$ with estimated standard deviation equal to 0.2801. This suggests that the mice in the germ-free environment developed lung tumor much earlier than those in the

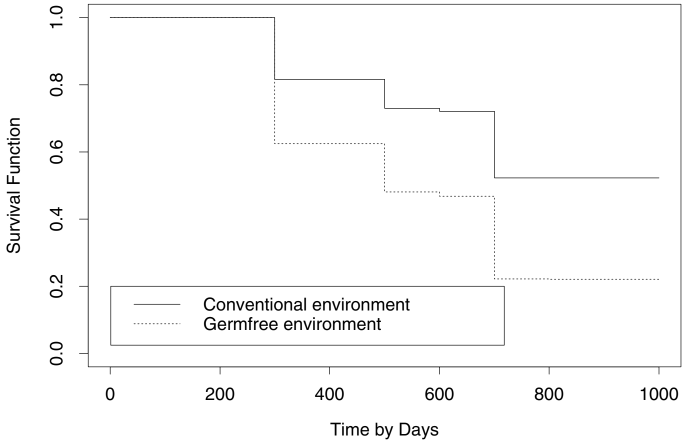

Fig. 5.4. Estimates of survival functions of tumor onset time.

conventional environment. To test this, the application of the score test for $\beta = 0$ discussed in the previous subsection yields $U_\beta(0, \hat{\boldsymbol{\alpha}}_0) = 6.1858$ and $\sigma^2(0, \hat{\boldsymbol{\alpha}}_0) = 8.5479$. This corresponds with a p-value of 0.034, and as before, suggests that the tumor occurrence rates of the two groups significantly differ. For estimation of the survival function, we take $s_j = 100j$ and define $\hat{S}(t; Z_i)$ as a step function with $\hat{S}(s_j; Z_i) = \prod_{k=1}^{j} \exp(-e^{\hat{\alpha}_k + Z_i \hat{\beta}})$. Figure 5.4 displays the estimated survival functions for animals in the two groups and they are similar to those given in Figure 5.1.

For the analysis above, one may be interested in how the number of intervals may affect the analysis. To study this, we repeated the analysis by partitioning $(0, 1000)$ into six intervals with the same length. That is, we have $m = 6$. In this case, the inference procedure produces $\hat{\beta} = 0.7795$ with estimated standard deviation being 0.3156, similar to those obtained above.

As another illustration, consider the data presented in Table 5.4 from a tumorigenicity experiment conducted in the Eppley Colony of University of Nebraska Medical Center. The data are reproduced from Ii et al. (1987) and Sun and Kalbfleisch (1993). For 100 male and 99 female rats and 15 intervals each of length 10 weeks, the numbers of animals that die within each of these intervals with or without tumors are recorded. Because, as seen from the table, there are no deaths within the first 3 intervals, it is natural to assume that there exist 12 time points, the end points of the other 12 intervals, at which the animals can die. That is, $m = 12$. The objective here is to compare the tumor occurrence rates between male and female rats.

Let T_i denote the tumor occurrence time and suppose that it can be described by the grouped PH model (1.10) with $Z_i = 0$ for male rats and 1 for female rats. Use of the maximum likelihood approach described in the previous subsection results in $\hat{\beta} = 0.7006$ with estimated standard error equal to 0.1969. The result indicates that the female rats seem to develop tumors significantly earlier than the male rats. For comparison, one can also use the score test given in the previous subsection. It produces $U_\beta(0, \hat{\boldsymbol{\alpha}}_0) = 14.3768$

Table 5.4. Observed numbers of rats that died with or without tumors within each of 15 10-week intervals

	Males		Females			Males		Females	
Weeks	Tumor	No	Tumor	No	Weeks	Tumor	No	Tumor	No
1-10					11-20				
21-30					31-40		3		
41-50		11	1	9	51-60			1	2
61-70	1	17	2	12	71-80		2		
81-90	1	3	2		91-100	4	5	5	1
101-110	10	5	7	5	111-120	8	8	18	
121-130	7		10	1	131-140	6	1	14	2
141-150	5	3	6	1					

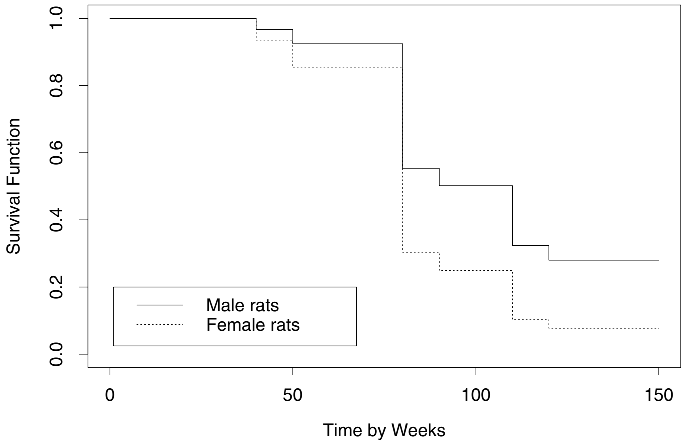

Fig. 5.5. Estimates of survival functions of tumor occurrence times.

and $\sigma^2(0, \hat{\boldsymbol{\alpha}}_0) = 27.1724$, giving a p-value of 0.006 and a similar conclusion. As in the previous example, we also calculated the estimated survival functions corresponding to the two groups and they are given in Figure 5.5. It is interesting to note from the figure that the tumor occurrence rates for male and female rats were similar during the earlier study weeks and the difference mainly occurred in later weeks.

5.5.3 Discussion

The inference approach discussed in this section is appropriate if the underlying survival time of interest is discrete or can be observed only at finite time points, which is often the case for periodic follow-up studies. The approach also applies to situations where the underlying survival time is continuous, but only finite discrete failure times are observable due to, for example, grouping. As most approaches developed for discrete models, the method given above is straightforward and simpler than those discussed in Sections 5.2 to 5.4 in theory because only finite-dimensional parameters are involved. On the other hand, it can be complicated in computation for large m and p, the dimension of regression parameters, because one has to deal with $I(\boldsymbol{\beta}, \boldsymbol{\alpha})$, which is a $(p + m) \times (p + m)$ matrix.

As an alternative to the grouped PH model, one can apply the logistic model (1.12) to regression analysis of discrete current status data. In this case, the log likelihood function has the form

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{j=1}^{m} \left\{ \sum_{i \in D_j} \log \left[\frac{1 - \prod_{k=1}^{j} (1 + \gamma_k e^{\boldsymbol{Z}'_i \boldsymbol{\beta}})^{-1}}{\prod_{k=1}^{j} (1 + \gamma_k e^{\boldsymbol{Z}'_i \boldsymbol{\beta}})^{-1}} \right] \right.$$

$$\left. - \sum_{i \in R_j} \sum_{k=1}^{j} \log(1 + \gamma_k e^{\boldsymbol{Z}'_i \boldsymbol{\beta}}) \right\}$$

in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_m)'$ defined in model (1.12). Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be carried out similarly as in Section 5.5.1 for the parameters in the grouped PH model.

5.6 Bibliography, Discussion, and Remarks

As remarked before, there exists extensive literature about current status data in the context of demographical studies (Diamond and McDonald, 1991; Diamond et al., 1986) and tumorigenicity experiments (Dinse and Lagakos, 1983; Dewanji and Kalbfleisch, 1986). In contrast, the literature about current status data arising from survival studies is limited, especially about regression analysis of current status data under the commonly used semiparametric, survival models. As discussed before, the articles that gave rigorous studies for the use of the PH model include Huang (1996) and Huang and Wellner (1997). In particular, they provided important and fundamental ground work for the asymptotic study of other similar inference procedures. Huang (1995) and Rossini and Tsiatis (1996) investigated the use of the proportional odds model, and the authors who discussed the additive hazards model for current status data include Ghosh (2001), Lin, Oakes, and Ying (1998), and Martinussen and Scheike (2002b).

Several other semiparametric models have also been considered for regression analysis of current status data in the literature. One is the accelerated failure time model (1.8), which provides a different way to describe the relationship between the survival time of interest and covariates. To fit this model to current status data, Shen (2000) developed a random sieve likelihood-based approach. Xue et al. (2004) investigated a partial linear model that is similar to model (1.8) and given by

$$\log T = \boldsymbol{Z}'\boldsymbol{\beta} + g(S) + W .$$

In this model, g is assumed to be an unknown smooth function, S represents some covariates that may have nonlinear effect on T, and W has a distribution known up to a scale parameter. For inference, they proposed to use the sieve maximum likelihood approach and in fact their method applies to more general response variable than just the log of survival time.

Also for the analysis of current status data, Sun and Sun (2005) studied the linear transformation model (1.9) and developed some estimating equation approaches for estimation of regression parameters. Shiboski (1998) proposed some generalized additive models and applied the maximum likelihood approach with the use of step function approximation for inference. Other models that have been investigated for regression analysis of current status data include the binary choice model (Huang and Wellner, 1996; Klein and Spady, 1993), generalized linear models (Jewell and Shiboski, 1990), and spline models (Grummer-Strawn, 1993).

There exist several other types of current status data that are not discussed here (Jewell and van der Laan, 1996) and for their analyses, inference approaches that take into account the special feature of each type of the data are needed. These include the current status data with time-dependent covariates (van der Laan and Robins, 1998), doubly censored current status data (Jewell and van der Laan, 1997, 2004a; Rabinowitz and Jewell, 1996; van der Laan and Andrews, 2000; van der Laan et al., 1997; van der Laan and Jewell, 2001), and case-cohort current status data (Jewell and van der Laan, 2004b; Shiboski and Jewell, 1992). The doubly censored current status data mean that the survival time of interest is defined as the elapsed time between two related events. In terms of the observed information, they imply that the initial event time is not observed, but with a known distribution and for the subsequent event, only current status data are available. A similar form of such data, doubly censored data, is considered in Chapter 8. Case-cohort current status data, by its name, refer to current status data arising from case-cohort studies. Also in practice, one may face current status data that arise from competing risk studies (Jewell et al., 2003), are generated from a cure survival model (Lam and Xue, 2005), or involve truncation in addition to censoring (Kim, 2003a).

In this chapter, three types of inference approaches are discussed for fitting current status data to various semiparametric regression models. In theory, the maximum likelihood approach applies to any model and is the most efficient method. As noted before, however, it may be a complicated approach both in terms of investigation of its properties and its implementation. Consequently, one may prefer the sieve maximum likelihood approach, the estimating equation approach, or other less efficient, but simpler approaches.

As in any regression analysis, a natural question for regression analysis of current status data is how to choose a better or an appropriate model among possible models or to assess the goodness-of-fit for a particular model. For this, of course, one and the first criterion is to use the prior knowledge, assuming that it exists, about the possible relationship between the survival variable of interest and covariates or how the covariates may affect the survival variable. In the case of right-censored failure time data, many statistical procedures and diagnosis tools have been proposed (Klein and Moeschberger, 2003; Lawless, 2003), but there exist few methods specifically developed for current status data. An exception is given by Ghosh (2003), who discussed the goodness-of-fit of the additive hazards model (5.9) and developed some numerical and graphical methods based on the inference approach described in Section 5.4. Babineau (2005) also provided some discussion about the goodness-of-fit for the fitting of parametric models to current status data. More discussion on this is given in Section 10.2.

This page intentionally blank

6

Regression Analysis of Case II Interval-censored Data

6.1 Introduction

This chapter discusses regression analysis of general or case II interval-censored failure time data. Compared with current status data, it is apparent that case II interval-censored data provide more information about the underlying survival time of interest. Thus intuitively, regression analysis of case II interval-censored data may seem to be simpler than that of current status data. On the other hand, for case II interval-censored data, one has to deal with two or more variables representing observation times rather than only one variable as in the case of current status data. As seen in Chapter 3 and the following, regression analysis of case II interval-censored data is more complicated and difficult than that of current status data in both computation and theory.

For general interval-censored data, as discussed before, three formulations or representations, (1.1), (1.2), and (1.3), can be used. The representation (1.1) is used more often in practice, and the representations (1.2) and (1.3), especially (1.2), are more convenient for formulating inference problems and investigating asymptotic properties of inference procedures. For a given set of interval-censored data, the three representations are equivalent from the likelihood point of view and one can easily transform the data given in (1.2) or (1.3) to (1.1). But the reverse transformation may not be obvious. For example, for a left- or right-censored observation given by (1.1), in terms of representation (1.2), only U or V is known by definition. From the likelihood point of view, however, one can take V (for left-censored observations) or U (for right-censored observations) to be any value in these situations because the corresponding term makes no contribution to the likelihood. This chapter discusses all three representations with Sections 6.2 and 6.5 mainly focusing on (1.1) and Sections 6.3 and 6.4 dealing with data described by (1.3).

For the analysis of case II interval-censored data, as in Chapter 5, we first discuss the use of the PH model (1.4) along with the maximum likelihood approach for inference in Section 6.2. Although the resulting inference procedure is similar to that described in Section 5.2, implementation and computation for case II interval-censored data are more complicated than for current status data. Also derivation of asymptotic properties is much harder. Section 6.3 considers regression analysis of case II interval-censored data using the proportional odds model and an approximate maximum likelihood approach is described for inference. The topic of Section 6.4 is use of the accelerated failure time model for analysis and for this model, an estimating equation approach is provided for inference about regression parameters. In Section 6.5, regression analysis of discrete case II interval-censored data is studied with use of the logistic model (1.12) and the maximum likelihood approach. Section 6.6 provides bibliographic notes about regression analysis of general interval-censored data along with some general remarks about approaches, problems, and issues in the analysis that are not treated in the preceding sections.

6.2 Analysis with the Proportional Hazards Model

Consider a survival study that consists of n independent subjects and gives rise to interval-censored data

$$\{\, (L_i, R_i], \boldsymbol{Z}_i \,;\, i \,=\, 1, ..., n \,\} \tag{6.1}$$

for the survival times of interest. Here as before, $(L_i, R_i]$ denotes the interval within which the survival event for the ith subject is observed to occur, and $\boldsymbol{Z}_i$ represents the p-dimensional vector of covariates from subject i, $i = 1, ..., n$. Also as before, let $S(t; \boldsymbol{Z})$ denote the survival function for a subject with covariates $\boldsymbol{Z}$. Then the likelihood function is proportional to

$$L \,=\, \prod_{i=1}^{n} \left[\, S(L_i, \boldsymbol{Z}_i) \,-\, S(R_i, \boldsymbol{Z}_i) \,\right]$$

assuming that $L_i < R_i$ for all $i = 1, ..., n$.

In this section, we assume that $S(t; \boldsymbol{Z})$ is specified by the PH model (1.4). The log of the likelihood function given above then has the form

$$l(\boldsymbol{\beta}, S_0) \,=\, \sum_{i=1}^{n} \log \left\{ \,[S_0(L_i)]^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} \,-\, [S_0(R_i)]^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} \, \right\}$$

in terms of the regression parameter $\boldsymbol{\beta}$ and the baseline survival function $S_0(t)$. For inference about $\boldsymbol{\beta}$ and S_0, we consider the maximum likelihood approach, first studied in Finkelstein (1986), and discuss some related asymptotic properties and survival comparison in the next two subsections. Two illustrative examples about the approach are then provided and followed by discussion of some other approaches for the inference.

6.2.1 Maximum Likelihood Estimation

This section discusses maximum likelihood estimation of $\boldsymbol{\beta}$ and S_0. As with the one-sample situation discussed in Sections 3.2 to 3.4, the likelihood depends on S_0 only through its values at the different observation time points. Thus one only needs to focus on estimating the values of S_0 at these time points. Let $s_0 = 0 < s_1 < ... < s_{m+1} = \infty$ denote the ordered distinct time points of all observed interval end points $\{L_i, R_i;\, i = 1, ..., n\}$ and $\alpha_{ij} = I(s_j \in (L_i, R_i])$, $j = 1, ..., m$, $i = 1, ..., n$. As in Section 5.2.1, suppose that $S_0(s_j)$ can be written as

$$S_0(s_j) \;=\; \prod_{k=1}^{j} e^{-\,\exp(\alpha_k)} \;=\; e^{-\sum_{k=1}^{j}\,\exp(\alpha_k)}\,,$$

$j = 1, ..., m$. Then in terms of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$, the log likelihood function $l(\boldsymbol{\beta}, S_0)$ can be rewritten as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{m+1} \alpha_{ij}\, \left[e^{-\,a_{j-1}\,\exp(\boldsymbol{Z}_i'\boldsymbol{\beta})} \;-\; e^{-\,a_j\,\exp(\boldsymbol{Z}_i'\boldsymbol{\beta})} \right] \right\}\,,$$

where $a_j = \sum_{k=0}^{j} \exp(\alpha_k)$, $\alpha_0 = -\infty$ and $\alpha_{m+1} = \infty$.

To maximize the log likelihood function above, we can treat it as a log likelihood function arising from a parametric model and use the Newton-Raphson algorithm as before. To this end, one needs the score functions of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and the observed Fisher information matrix. The score functions are

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \;=\; \sum_{i=1}^{n} \boldsymbol{Z}_i\, g_i^{-1} \sum_{j=1}^{m+1} \alpha_{ij}\, (f_{i\,j-1} \,-\, f_{ij})$$

and

$$U_{\alpha_j}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} \;=\; \sum_{i=1}^{n} g_i^{-1}\, b_{ij}\, c_{ij}\,,$$

where $f_{ij} = S(s_j; \boldsymbol{Z}_i)\, \log S(s_j; \boldsymbol{Z}_i)$, $f_{i\,0} = f_{i\,m+1} = 0$, $b_{ij} = \exp(\alpha_j + \boldsymbol{Z}_i'\boldsymbol{\beta})$, $c_{ij} = \sum_{l=j}^{m+1} (\alpha_{il} \,-\, \alpha_{i\,l+1})\, S_{il}(s_l; \boldsymbol{Z}_i)$, $\alpha_{i\,m+2} = 0$, and

$$g_i \;=\; \sum_{j=1}^{m+1} \alpha_{ij}\, \left[S(s_{j-1}; \boldsymbol{Z}_i) \,-\, S(s_j; \boldsymbol{Z}_i) \right]\,,$$

$j = 1, ..., m$, $i = 1, ..., n$. Then the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be determined by solving the score equations

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; 0\;,\; U_{\alpha_j}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; 0\;,\; j \;=\; 1, ..., m\,.$$

Let

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \begin{pmatrix} I_{11} \; I_{12} \\ I_{21} \; I_{22} \end{pmatrix}$$

denote the observed Fisher information matrix, where

$$I_{11} \;=\; -\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \,, \, I_{12} \;=\; I_{21}' \;=\; -\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \boldsymbol{\beta}} \right) \,, \, I_{22} \;=\; -\left(\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} \right) \,.$$

Then we have

$$I_{11} = \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i' \left\{ \left[\frac{\sum_{j=1}^{m+1} \alpha_{ij} (f_{i\,j-1} - f_{ij})}{g_i} \right]^2 - \frac{\sum_{j=1}^{m+1} \alpha_{ij} (h_{i\,j-1} - h_{ij})}{g_i} \right\} ,$$

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{Z}_i b_{ij} \left[\frac{c_{ij} + \sum_{l=j}^{m+1} (\alpha_{i\,l} - \alpha_{i\,l+1}) S(s_l; \boldsymbol{Z}_i) \log S(s_l; \boldsymbol{Z}_i)}{g_i} \right.$$

$$\left. - \frac{c_{ij}}{g_i^2} \sum_{l=1}^{m+1} \alpha_{il} (f_{i\,l-1} - f_{il}) \right] ,$$

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j^2} = \sum_{i=1}^{n} b_{ij} \, c_{ij} \left[\frac{1 - b_{ij}}{g_i} - \frac{b_{ij} \, c_{ij}}{g_i^2} \right] ,$$

and

$$\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} = -\sum_{i=1}^{n} \left(\frac{b_{ij} \, b_{ik} \, c_{ij} \, c_{ik}}{g_i^2} + \frac{b_{ij} \, b_{ik} \, c_{ik}}{g_i} \right) \text{ for } j < k \,,$$

where

$$h_{ij} \;=\; f_{ij} \log S(s_j; \boldsymbol{Z}_i) \;+\; f_{ij} \,,$$

and $h_{i0} \;=\; h_{i\,m+1} \;=\; 0$.

For implementation of the Newton-Raphson algorithm, the remarks given in Section 5.2.1 apply. In particular, using the form of the inverse of a symmetric 2 × 2 partition matrix, we have

$$I^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \begin{pmatrix} I_{11|2}^{-1} \; I_{12|2} \\ I_{21|1} \; I_{22|1} \end{pmatrix} ,$$

where $I_{11|2} \;=\; I_{11} \,-\, I_{12} I_{22}^{-1} I_{21}$, $I_{12|2} \;=\; I_{21|1}' \;=\; -I_{11|2}^{-1} I_{12} I_{22}^{-1}$ and $I_{22|1} \;=\; I_{22}^{-1} \,+\, I_{22}^{-1} I_{21} I_{11|2}^{-1} I_{12} I_{22}^{-1}$.

6.2.2 Asymptotic Properties and Survival Comparisons

Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\alpha}}_n \;=\; (\hat{\alpha}_1, ..., \hat{\alpha}_m)'$ denote the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ defined in the previous subsection for given n. As in Section 5.2.1, define $\hat{S}_n(t)$, the estimator of the baseline survival function $S_0(t)$, to be the right-continuous step function with jumps only at the s_j's and

$$\hat{S}_n(s_j) = \prod_{k=1}^{j} e^{-\exp(\hat{\alpha}_k)},$$

$j = 1, ..., m$. Also define $\hat{\Lambda}_n(t) = -\log \hat{S}_n(t)$, an estimator of the baseline cumulative hazard function $\Lambda_0(t)$. Assume that $S_0(t)$ is continuous and the $\boldsymbol{Z}_i$'s are bounded. To describe the conditions required for the asymptotic properties of $\hat{\boldsymbol{\beta}}_n$ and $\hat{S}_n(t)$, suppose that the observed data are given by the representation (1.2). That is, the observed data are given in terms of U and V. Then under the same conditions as those for current status data, it can be shown that both $\hat{\boldsymbol{\beta}}_n$ and $\hat{S}_n(t)$ are consistent (Huang and Wellner, 1997). In particular, like the estimator of the baseline survival function given in Section 5.2 based on current status data, we have the following results: if both U and V are discrete, then

$$\hat{S}_n(t) \to S_0(t)$$

almost surely at all the mass points of U and V; if at least one of U and V has a continuous distribution function, then

$$\sup_{0 \le t < \infty} |\, \hat{S}_n(t) - S_0(t) \,| \to 0$$

almost surely.

For the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$, as in the case of current status data, we need more conditions. These include that (a) the union of the support of U and V is contained in a bounded interval that is bounded away from zero and (b) S_0 has strictly positive and bounded continuous derivative on the support interval defined in (a). Also assume that condition (I3) in Section 3.5.2 holds. That is, no exact failure time is observed. Then under some regularity conditions, as $n \to \infty$, one has that

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to N(0, \Sigma^{-1})$$

in distribution and Σ^{-1} can be estimated by $I_{11|2}^{-1}$ given in the previous subsection with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ replaced by their maximum likelihood estimators (Huang and Wellner, 1997). In the expression above, Σ denotes the information lower bound for $\boldsymbol{\beta}$ and thus $\hat{\boldsymbol{\beta}}_n$ is asymptotically efficient.

As discussed in Chapter 4, the comparison of several survival functions is often of interest in practice. Under the model considered here, if one takes $\boldsymbol{Z}_i$ to be the group indicator vector for subject i, the comparison is then equivalent to testing $\boldsymbol{\beta} = 0$, which can be performed naturally using the score test. One main advantage of the score test is that it only involves the maximum likelihood estimator denoted by $\hat{\boldsymbol{\alpha}}_0$ of $\boldsymbol{\alpha}$ at $\boldsymbol{\beta} = 0$, but not the maximum likelihood estimator of $\boldsymbol{\beta}$. This can save a great deal of computational effort compared with the Wald test based on $\hat{\boldsymbol{\beta}}$. The score test statistic, defined as $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_0$, has the form

$$U_{PH} = \sum_{i=1}^{n} \boldsymbol{Z}_i \frac{\sum_{j=1}^{m+1} \alpha_{ij} \left[\hat{S}_0(s_{j-1}) \log \hat{S}_0(s_{j-1}) - \hat{S}_0(s_j) \log \hat{S}_0(s_j) \right]}{\sum_{j=1}^{m+1} \alpha_{ij} \left[\hat{S}_0(s_{j-1}) - \hat{S}_0(s_j) \right]} , \quad (6.2)$$

where $\hat{S}_0(t) = \hat{S}_n(t)$ at $\boldsymbol{\beta} = 0$. Finkelstein (1986) first discussed this score test and suggested approximating the variance-covariance matrix of U_{PH} by $I_{11|2}$ and thus the distribution of $U'_{PH} \, I_{11|2}^{-1} \, U_{PH}$ by the χ^2 distribution with the degrees of freedom p.

Implementation of the score test requires determination of $\hat{\boldsymbol{\alpha}}_0$ or $\hat{S}_0(t)$, which is basically the one-sample estimation problem discussed in Sections 3.3 and 3.4. Thus compared with $\hat{\boldsymbol{\alpha}}$ or $\hat{S}_n$, the restricted estimator of $\boldsymbol{\alpha}$ or S_0 can be easily obtained by, for example, the self-consistency algorithm described in Section 3.4.1. Because the score statistic U_{PH} is a function of $\hat{S}_0(t)$, it is convenient to directly estimate the baseline survival function S_0 at $\boldsymbol{\beta} = 0$. Using the self-consistency algorithm, we have the restricted maximum likelihood estimator of S_0 given by the following self-consistency equations

$$\hat{S}_0(s_j) \; = \; \hat{S}_0(s_{j-1}) \left(1 \; - \; \frac{\sum_{i=1}^{n} \hat{g}_i^{-1} \alpha_{ij} [\hat{S}_0(s_{j-1}) - \hat{S}_0(s_j)]}{\sum_{k=j}^{m+1} \sum_{i=1}^{n} \hat{g}_i^{-1} \alpha_{ij} [\hat{S}_0(s_{k-1}) - \hat{S}_0(s_k)]} \right) ,$$

$j \; = \; 1, ..., m$, where $\hat{S}_0(s_0) \; = \; 1$ and

$$\hat{g}_i \; = \; \sum_{j=1}^{m+1} \alpha_{ij} \left[\hat{S}_0(s_{j-1}) - \hat{S}_0(s_j) \right] ,$$

$i \; = \; 1, ..., n$.

6.2.3 Two Examples

In this subsection, we discuss two illustrative examples for the inference procedure described in the previous subsections. The first example deals with the breast cancer data presented in Table 1.4 and studied in Sections 2.3.4, 2.4.3, 3.4.4, and 4.5.2, and the second example concerns the hemophilia data given in data set II of Appendix A and discussed in Section 3.4.4.

For the breast cancer data, as before, define T_i to be the time to breast retraction for patient i, $i \; = \; 1, ..., 94$. Also define $Z_i \; = \; 0$ for the patients given the radiation therapy only (RT) and 1 for those given the radiation therapy plus adjuvant chemotherapy (RCT). As discussed in Section 4.5.2, for this data set, it seems reasonable to assume that observation times are independent of the survival times T_i's. Also the result presented there indicates that the patients in the two treatment groups had significantly different breast retraction rates.

Suppose that the T_i's can be reasonably described by the PH model (1.4). The application of the inference procedure described in the previous subsections gives $\hat{\beta}_n \; = \; 0.8002$ with estimated standard deviation equal to 0.290.

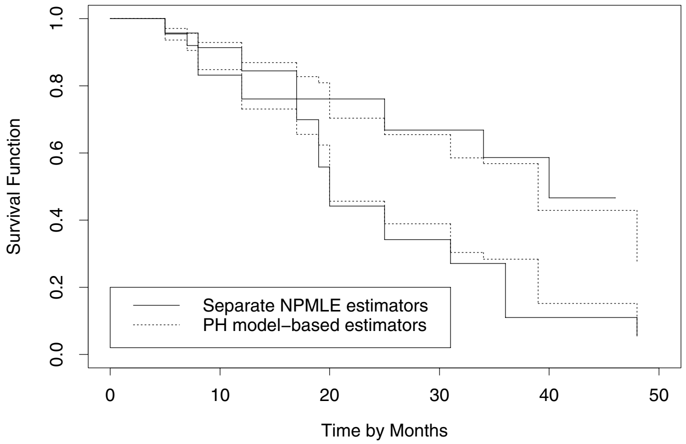

Fig. 6.1. Estimates of survival functions of time to breast retraction: top, RT; bottom, RCT.

Based on the standard normal distribution, this yields a p-value of 0.006 for testing $\beta = 0$. Using the score test for the comparison of the two groups, we obtain $U_{PH} = 10.0006$ and a p-value of 0.005. As expected, these results are similar to those given in Section 4.5.2 based on rank-based approaches. Figure 6.1 presents the maximum likelihood estimators of the survival functions for patients in the two groups. For comparison and model-checking, we include in the figure the separate NPMLEs of the same survival functions shown in Figure 3.2 that were obtained by the self-consistency algorithm. Figure 6.1 suggests that the PH model seems to provide a reasonable fit to the data set.

For the hemophilia data, as in Section 3.4.4, define T_i to be the time to HIV-infection for patient i, $i = 1, ..., 368$. In this data set, as with the breast cancer data, we have one covariate representing treatment group. One main objective of the study is to compare the HIV-infection risks between the patients who received no factor VIII concentrate and those who received up to 20,000 U factor VIII concentrate for their treatment. Define $Z_i = 0$ for the patients in the no dose group and 1 otherwise and assume that the T_i's follow the PH model (1.4). Using the maximum likelihood approach described in the previous subsections, we obtain $\hat{\beta}_n = 1.8644$ with an estimated standard error of 0.221. Based on the standard normal distribution, the p-value for testing $\beta = 0$ is close to zero. The same conclusion is given by the score test based on U_{PH}. These results indicate that the factor VIII blood concentrate given to the patients significantly increased their HIV-1 infection risks.

As with Figure 6.1, Figure 6.2 displays the estimated survival functions of the time to HIV-1 infection obtained separately and under the PH model,

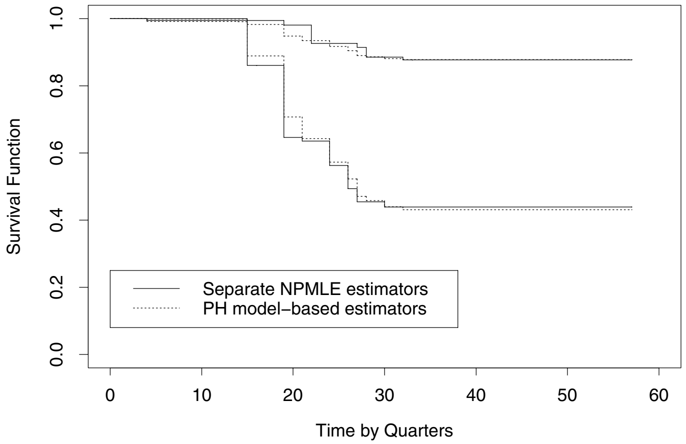

Fig. 6.2. Estimates of survival functions of time to HIV-1 infection: top, No factor VIII; bottom, Low dose VIII.

respectively, for the patients in the two groups. It indicates that the PH model fits the hemophilia data extremely well. The figure also confirms the test results given above and shows that the difference between the HIV-1 infection risks started at about 16 quarters.

6.2.4 Other Approaches

In addition to the full likelihood approach discussed in the previous subsections, several other approaches are available for fitting the PH model to general interval-censored failure time data. One is the marginal likelihood approach based on the likelihood given by the sum over all rankings of the underlying and unobserved failure times that are consistent with the observed censoring intervals. This approach is a direct generalization of the corresponding approach for right-censored failure time data (Kalbfleisch and Prentice, 1973) and reduces to that based on the partial likelihood when right-censored data are available. One main advantage of the method is that it does not require estimation of the baseline cumulative hazard function. On the other hand, the approach requires solving complicated score equations and involves a great deal of computational effort. Satten (1996) investigated this approach and proposed a Gibbs sampling procedure for generating underlying rankings and the use of stochastic approximation for solving the score functions. Goggins et al. (1998) also studied the approach and developed a Monte Carlo EM algorithm for determination of regression parameter estimates.

It is seen that the full likelihood approach directly estimates the finite-dimensional regression parameters and the infinite-dimensional nuisance parameter simultaneously. In contrast, the marginal likelihood approach focuses only on the finite-dimensional regression parameters. An approach that lies between the two approximates the infinite-dimensional nuisance parameter using some smooth, finite-dimensional parameters. Betensky et al. (2002) considered such an approach that applies the local likelihood described in Section 3.5.3 to fit the PH model to interval-censored data. For the same problem, Cai and Betensky (2003) proposed a penalized spline-based approach. In both approaches, some finite-dimensional functions are used to approximate the log baseline hazard function in the PH model.

6.3 Analysis with the Proportional Odds Model

This section deals with the proportional odds model (1.5) and assumes that the observed interval-censored failure time data are given in the form

$$\{\, K_i,\, (\, U_{ij},\, \delta_{ij} = I(U_{ij-1} < T_i \leq U_{ij})\,)_{j=1}^{K_i},\, \boldsymbol{Z}_i\, ;\, i \,=\, 1, ..., n\,\}\, .$$

In this expression, for subject i, $\boldsymbol{Z}_i$ denotes the vector of covariates, K_i the number of observations and $U_{i1} \,<\, ... \,<\, U_{iK_i}$ the observation time points with $U_{i0} \,=\, 0$. That is, the study consists of n independent subjects and each subject is observed at a sequence of time points that are assumed to be independent of the survival time of interest for the subject, T_i. The likelihood function then has the form

$$L(\boldsymbol{\beta}, H) = \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left\{ \frac{1}{1 + \exp[H(U_{ij-1}) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} - \frac{1}{1 + \exp[H(U_{ij}) + \boldsymbol{Z}_i'\boldsymbol{\beta}]} \right\}^{\delta_{ij}} ,$$

where $H(t) \,=\, -\text{logit}[S_0(t)]$ as before.

As in the previous section, it is natural to directly maximize the likelihood function above with respect to $\boldsymbol{\beta}$ and H and its implementation is similar to that of the approach in Section 6.2. Instead, we present an approximate conditional likelihood approach first investigated by Rabinowitz et al. (2000), which is much simpler than the maximum likelihood approach. The resulting approximate conditional likelihood leaves H or S_0 arbitrary and only involves the regression parameter $\boldsymbol{\beta}$. Here it is assumed that one is interested only in $\boldsymbol{\beta}$ with H or S_0 being a nuisance parameter.

6.3.1 An Approximate Conditional Likelihood Approach

Let Ω denote the set of all observation time points $\{\, U_{ij}\, ;\, j \,=\, 1, ..., K_i,\, i \,=\, 1, ..., n\,\}$ and $\{\Omega_l\, ;\, l \,=\, 1, ..., k\}$ a partition of Ω such that all U_{ij} in each Ω_l are close to each other. Define $Y_{ij} \,=\, I(T_i \,\leq\, U_{ij}) \,=\, \sum_{l=1}^{j} \delta_{il}$, indicating if the survival event of interest for subject i has occurred before or at U_{ij}, $j = 1, ..., K_i$, $i = 1, ..., n$.

To derive the approximate conditional likelihood, for each l, consider all Y_{ij} for which the corresponding observation time points U_{ij} belong to Ω_l and assume that they are from different subjects. That is, they are independent. Then one has

$$P(\, Y_{ij} \,|\, U_{ij} \in \Omega_l \,,\, \boldsymbol{Z}_i \,) \;=\; \prod_{U_{ij} \in \Omega_l} \frac{\exp\{[H(U_{ij}) + \boldsymbol{Z}_i'\boldsymbol{\beta}]\, Y_{ij}\}}{1 \,+\, \exp\{H(U_{ij}) + \boldsymbol{Z}_i'\boldsymbol{\beta}\}}$$

and the conditional likelihood

$$L_l^*(\boldsymbol{\beta}, H) \;=\; P(\, Y_{ij} \,|\, y \,,\, U_{ij} \in \Omega_l \,,\, \boldsymbol{Z}_i \,)$$

$$= \frac{P(\, Y_{ij} | U_{ij} \in \Omega_l,\, \boldsymbol{Z}_i \,)}{P(\, Y_{ij} = y_{ij} \,,\, \sum_{U_{ij} \in \Omega_l} y_{ij} \;=\; y \,|\, U_{ij} \in \Omega_l,\, \boldsymbol{Z}_i \,)}$$

$$= \frac{\prod_{U_{ij} \in \Omega_l} \exp\{[H(U_{ij}) + \boldsymbol{Z}_i'\boldsymbol{\beta}]\, Y_{ij}\}}{\sum_{(y)} \prod_{U_{ij} \in \Omega_l} \exp\{[H(U_{ij}) + \boldsymbol{Z}_i'\boldsymbol{\beta}]\, y_{ij}\}}$$

given $y \;=\; \sum_{U_{ij} \in \Omega_l} Y_{ij}$, where $\sum_{(y)}$ denotes the summation over all permutations of the y_{ij}'s whose corresponding U_{ij} belong to Ω_l.

In $L_l^*(\boldsymbol{\beta}, H)$, if all U_{ij} in Ω_l are identical, then it reduces to

$$L_l(\boldsymbol{\beta}) \;=\; L_l^*(\boldsymbol{\beta}, H) \;=\; \frac{\prod_{U_{ij} \in \Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, Y_{ij})}{\sum_{(y)} \prod_{U_{ij} \in \Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij})} \,.$$

That is, $L_l^*(\boldsymbol{\beta}, H)$ is independent of H and involves $\boldsymbol{\beta}$ only. In practice, of course, this may not be the case. On the other hand, all time points in Ω_l should be close to each other by definition. This suggests that one can use the approximate conditional likelihood

$$L_c(\boldsymbol{\beta}) \;=\; \prod_{l=1}^{k} L_l(\boldsymbol{\beta}) \;=\; \prod_{l=1}^{k} \frac{\prod_{U_{ij} \in \Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, Y_{ij})}{\sum_{(y)} \prod_{U_{ij} \in \Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij})}$$

for inference about $\boldsymbol{\beta}$.

It is seen right away that one major advantage of the approximate conditional likelihood L_c is that it is free of the baseline survival function S_0 or H. Also it can be easily seen that L_c has the same format as the partial likelihood arising from the logistic regression (Lawless, 2003). Thus one can apply the standard statistical software for logistic regression of right-censored failure time data to maximize L_c by regarding $\{Y_{ij}\,;\, j \;=\; 1, ..., K_i\}$ to be generated from independent subjects. One way to take into account their dependence, suggested by Rabinowitz et al. (2000), is to treat the resulting score function of $\boldsymbol{\beta}$ as an estimating equation in estimating the variance-covariance matrix of the resulting estimators.

Specifically, let $\hat{\boldsymbol{\beta}}_c$ denote the estimator of $\boldsymbol{\beta}$ given by the value of $\boldsymbol{\beta}$ that maximizes L_c. Rabinowitz et al. (2000) argue that $\hat{\boldsymbol{\beta}}_c$ is consistent under some regularity conditions. Define

$$E_l(\boldsymbol{\beta}) \,=\, \frac{\sum_{(y)} \left(\sum_{U_{ij}\in\Omega_l} \boldsymbol{Z}_i\, y_{ij}\right) \prod_{U_{ij}\in\Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij})}{\sum_{(y)} \prod_{U_{ij}\in\Omega_l} \exp(\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij})}\,,$$

$l \,=\, 1, ..., k$. Then the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_c$ can be estimated by $\hat{\Sigma}_{\boldsymbol{\beta}} \,=\, A^{-1}(\hat{\boldsymbol{\beta}}_c)\, B(\hat{\boldsymbol{\beta}}_c)\, A^{-1}(\hat{\boldsymbol{\beta}}_c)$, where

$$A(\boldsymbol{\beta}) \,=\, -\frac{\partial^2 \log L_C(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}$$

$$= \sum_{l=1}^{k} \left[\frac{\sum_{(y)} \left(\sum_{U_{ij}\in\Omega_l} \boldsymbol{Z}_i\, y_{ij}\right)^{\otimes 2} \prod_{U_{ij}\in\Omega_l} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij}}}{\sum_{(y)} \prod_{U_{ij}\in\Omega_l} e^{\boldsymbol{Z}_i'\boldsymbol{\beta}\, y_{ij}}} - E_l(\boldsymbol{\beta}) E_l'(\boldsymbol{\beta}) \right]$$

and

$$B(\boldsymbol{\beta}) \,=\, \sum_{l=1}^{k} \sum_{l'} \left[\sum_{U_{ij}\in\Omega_l} \boldsymbol{Z}_i\, Y_{ij} - E_l(\boldsymbol{\beta}) \right] \left[\sum_{U_{i'j'}\in\Omega_{l'}} \boldsymbol{Z}_{i'}\, Y_{i'j'} - E_{l'}(\boldsymbol{\beta}) \right]' \,.$$

In the above, $\boldsymbol{a}^{\otimes 2} \,=\, \boldsymbol{a}\,\boldsymbol{a}'$ for a vector $\boldsymbol{a}$ and for each l, $\sum_{l'}$ denotes the summation over all $l' \,=\, 1, ..., k$ such that there exist $U_{ij} \,\in\, \Omega_l$ and $U_{ij'} \,\in\, \Omega_{l'}$ for some i. For such Ω_l and $\Omega_{l'}$ with $l \neq l'$, we say that they are correlated.

It is easy to see from the derivation of the approximate conditional likelihood that in the case of current status data, L_c is indeed a conditional likelihood function because $K_i \,=\, 1$. In this case, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_c$ can be estimated by $A^{-1}(\hat{\boldsymbol{\beta}}_c)$, the inverse of the observed information matrix. The standard statistical software for logistic regression of right-censored data gives $A^{-1}(\hat{\boldsymbol{\beta}}_c)$ as the covariance estimate of $\hat{\boldsymbol{\beta}}_c$ for general situations.

As discussed in Section 6.2.2, the comparison of several survival functions is often of interest and can be performed using the score test for testing $\boldsymbol{\beta} \,=\, 0$ with $\boldsymbol{Z}_i$ taken to be the group indicator. For the current situation, the score test statistic has the form

$$U_{PO} \,=\, \frac{\partial \log L_c(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \Bigg|_{\boldsymbol{\beta}=0} \,=\, \sum_{l=1}^{k} \left[\sum_{U_{ij}\in\Omega_l} \boldsymbol{Z}_i\, Y_{ij} - E_l(\boldsymbol{0}) \right] \,,$$

where

$$E_l(\boldsymbol{0}) \,=\, |y|^{-1} \sum_{(y)} \left(\sum_{U_{ij}\in\Omega_l} \boldsymbol{Z}_i\, y_{ij} \right)$$

with $|y|$ denoting the number of terms in the summation $\sum_{(y)}$. It is easy to see that the statistic U_{PO} is similar to statistics U_{cr} and U_r defined in Sections 4.2 and 4.3, respectively. All are summations of the differences between the observed and expected numbers of survival events. Under $\boldsymbol{\beta} = 0$, one can estimate the variance-covariance of U_{PO} by $B(\mathbf{0}) = B(\boldsymbol{\beta} = 0)$ and approximate the distribution of $U'_{PO}\, B^{-1}(\mathbf{0})\, U_{PO}$ by the χ^2 distribution with the degrees of freedom p.

To apply the approximate conditional likelihood approach, one needs to choose the partition $\{ \Omega_l \}$, which serves as comparison or risk sets. For this, it is easy to see that a better partition is one in which not too many subsets Ω_l's are correlated with each other and the ideal one has no two subsets that are correlated. In the latter case, L_c becomes a true conditional likelihood. One general approach that can be used to control the number of correlated subsets is to keep the number of elements or size of each subset Ω_l small. In addition to increasing the number of correlated Ω_l's, large sets can also significantly increase the computational effort required. To choose $\{ \Omega_l \}$, a simple approach, referred to as approach I below, is to put the same observation times from different subjects into one risk set Ω_l assuming that at each observation time, more than one subject is observed. Approach II, suggested by Rabinowitz et al. (2000), divides Ω evenly into subsets of a given size, assuming that the time points in Ω are ordered from the smallest to the largest. More comments on this are provided in the next subsection through an example.

The simulation studies conducted by Rabinowitz et al. (2000) indicate that both the estimation of $\boldsymbol{\beta}$ and its variance as well as the normal approximation to the distribution of $\hat{\boldsymbol{\beta}}_c$ seem to work well for situations with possibly small $\boldsymbol{\beta}$ or larger sample sizes. But for situations where $\boldsymbol{\beta}$ may be large and the sample size is small, some care is needed because the procedure may infrequently produce some very large $\hat{\boldsymbol{\beta}}_c$. In these cases, one could use larger Ω_l to reduce the degree of this problem, but as discussed above, it is apparent that the size of Ω_l should not be too large.

6.3.2 An Example

Consider data set III of Appendix A arising from an AIDS clinical trial, ACTG 359, that was designed to compare 6 different antiretroviral treatments regimens for AIDS patients (Gómez et al., 2003; Gulick et al., 2000). In AIDS studies, the viral load level of patients is often of interest and commonly measured by the number of RNA copies. Among others, one variable of interest is the first time at which the number of RNA copies drops below the threshold of 500 viral copies/ml, an event that is often used to indicate an AIDS patient's stage. In reality, viral load is usually only monitored periodically, meaning that the occurrence time of the event is interval-censored instead of being known exactly.

For the study, the blood samples of patients were supposed to be collected before the study and at month 1, 2, 3, 4, 6, 8, 10, and 12 for the determination of their RNA copy counts. As expected, many patients dropped out of the study or stopped providing blood samples earlier. Also some missed their scheduled visits. For patient i, let T_i denote the first time at which the number of RNA copies of the patient drops below 500 viral copies/ml during the study. Data set III of Appendix A gives the observed information about the observation process and the T_i's from 271 AIDS patients in ACTG 359 whose numbers of RNA copies were measured at least once during the 12 months period in addition to their initial numbers of RNA copies. Specifically, for each patient, the data set indicates if the blood sample was collected at each of 8 observation time points (a dot means no observation) and gives Y_{ij} defined in the previous subsection, $i = 1, ..., 271$, $j = 1, ..., 8$. Here Y_{ij} indicates (by 1) if the RNA count of subject i has already dropped below 500 by the jth observation time. All patients had initial RNA counts over 500 viral copies/ml. The data set consists of two parts. The first part, group 1, is for the patients whose initial numbers of RNA copies were below 20000 viral copies/ml and the second part, group 2, is for those whose initial numbers of RNA copies were above 20000 viral copies/ml, another threshold often used to indicate the stage of AIDS patients. The goal here is to assess if the initial number of RNA copies has prognostic effect on the time at which the RNA count drops below 500 copies/ml.

Define $Z_i = 0$ for the subjects in group 1 and 1 otherwise and assume that the T_i's follow the proportional odds model (1.5). For estimation of the effect of the initial RNA count on the T_i's using the approach given in the previous subsection, we first need to select a partition $\{ \Omega_l \}$. For this, it is apparent that we should not use approach I because it would divide Ω into $k = 8$ subsets and each would contain too many observations. The direct use of approach II would generate a partition in which many subsets are correlated. Hence, we choose to apply the two approaches together as follows. First we randomly divide the 271 patients evenly into small groups of size r and then apply approach I to each group. For example, for groupings of size 6, we have 45 small groups, and each group consists of 6 subjects except one group has 7 subjects. For the generated partition, each subset Ω_l is correlated with at most 7 other subsets. Table 6.1 gives the estimated effect $\hat{\beta}_c$ of the initial RNA count for several values of r, group size, and their corresponding estimated standard errors. It can be seen that the results seem to be reasonably stable for different r and all resulting p-values for testing $\beta = 0$ are close to zero. The

Table 6.1. Estimates of regression parameter in the proportional odds model for RNA data

Group size r	$\hat{\beta}_c$	SD of $\hat{\beta}_c$	Group size r	$\hat{\beta}_c$	SD of $\hat{\beta}_c$
4	-1.6428	0.2989	5	-1.8064	0.3049
6	-1.7830	0.3152	7	-1.6249	0.2912
8	-1.7447	0.2712	9	-1.6951	0.2763

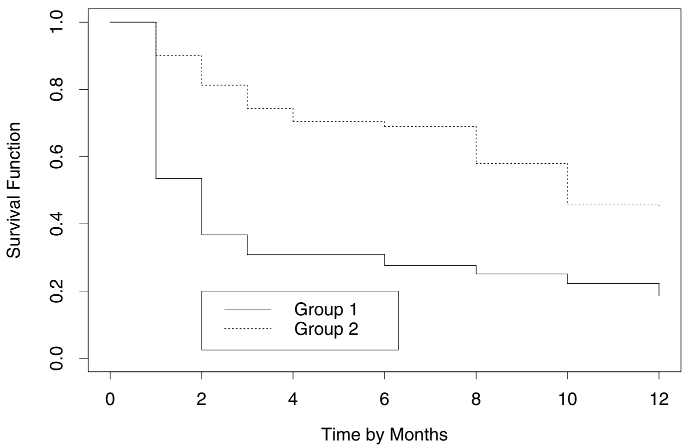

Fig. 6.3. NPMLEs of survival functions for RNA data.

results suggest that the initial RNA count has a very significant prognostic effect on the T_i's as expected, and the patients with lower initial RNA counts have their RNA counts dropping below 500 much earlier than those with higher initial RNA counts.

For the comparison of the two groups, one can also apply the score test based on U_{PO} and it gives similar results. For example, with $r = 6$, we obtain $U_{PO} = -95.7810$ with p-value equal to almost zero for testing $\beta = 0$ based on the χ^2 distribution with degree of freedom 1. To give a graphical comparison, Figure 6.3 presents the NPMLEs of the survival functions of T_i for the patients in the two groups and confirms the results given above. It can also be seen from the figure that for the patients in group 1, about 81% of them had their RNA counts reach 500 copies by 12 months, whereas the corresponding percentage for the patients in group 2 is only about 55%. We remark that for the partition used above with small r, some subsets Ω_l could end up with one or two elements. However, the results given in Table 6.1 suggest that this does not seem to have much effect on the analysis for the situation considered here.

6.3.3 Discussion

The approximate conditional likelihood approach applies to both current status data and the interval-censored data given in the form (1.2), but not the interval-censored data given in the form (1.1). For the data in the form (1.1), it is apparent that the observation times L and R and the survival time of interest are not independent. For the data given in the form (1.3), the approach requires that even after the survival event, the observation on study subject continues with the observation times recorded. This is needed to ensure the independence between observation times and the survival time of interest given covariates.

As commented above, a major advantage of the approximate conditional likelihood approach is that one does not need to estimate the baseline survival function, which is an infinite-dimensional nuisance parameter. Also, the idea behind it is straightforward, and one can make use of the standard statistical software for logistic regression to determine estimates of the regression parameters and their covariance. A drawback of the approach is that determination of the variance-covariance estimator, specifically $B(\hat{\boldsymbol{\beta}}_C)$, can become too complicated computationally in some situations, for example, one in which every subject contributes an observation time to each Ω_l. Furthermore, no rigorous investigation of the asymptotic properties of the approach has been conducted.

Other authors who considered regression analysis of general interval-censored failure time data using the proportional odds model include Huang and Rossini (1997), Huang and Wellner (1997), and Shen (1998). In particular, Huang and Wellner (1997) discussed the maximum likelihood approach, gave the efficient score function for $\boldsymbol{\beta}$, and established the asymptotic normal distribution of the maximum likelihood estimator of $\boldsymbol{\beta}$. Huang and Rossini (1997) and Shen (1998) considered the sieve maximum likelihood approach with using the linear functions (5.8) and monotone spline functions, respectively, to approximate the nuisance function.

6.4 Analysis with the Accelerated Failure Time Model

In this section, we consider the same scenario as in the previous section. Specifically, we assume that each study subject is observed at a sequence of time points with the data given by

$$\{\, K_i,\; (\, U_{ij},\; \delta_{ij} = I(U_{ij-1} < T_i \le U_{ij})\,)_{j=1}^{K_i},\; \boldsymbol{Z}_i\,;\; i \;=\; 1, ..., n\,\}\,,$$

which is the same notation used in the previous section. Instead of the proportional odds model, it is assumed that the survival times of interest, T_i, follow the accelerated failure time model, that is, the log linear model (1.8). Also it is assumed that the U_{ij}'s and T_i are independent given $\boldsymbol{Z}_i$ and one is interested only in the regression parameter $\boldsymbol{\beta}$.

Let F denote the distribution of W in model (1.8) and for a p-dimensional vector $\boldsymbol{b}$, define $U_{ij}(\boldsymbol{b}) \;=\; \log(U_{ij}) \;-\; \boldsymbol{Z}_i'\boldsymbol{b}$, $j \;=\; 1, ..., K_i$, $i \;=\; 1, ..., n$. Also define $Y_{ij} \;=\; I(T_i \le U_{ij})$ as in the previous section and for each i, let U_{iL} and U_{iR} denote the two U_{ij} that are the last observation time for which $Y_{ij} \;=\; 0$ and the first observation time for which $Y_{ij} \;=\; 1$, respectively. That is, $(U_{iL}, U_{iR}]$ is the observed interval to which T_i belongs. The likelihood function is then proportional to

$$L(\boldsymbol{\beta}, F) = \prod_{i=1}^{n} \left[F(U_{iR}(\boldsymbol{\beta})) - F(U_{iL}(\boldsymbol{\beta})) \right] , \tag{6.3}$$

where $U_{iL}(\boldsymbol{b}) = \log(U_{iL}) - \boldsymbol{Z}_i' \boldsymbol{b}$ and $U_{iR}(\boldsymbol{b}) = \log(U_{iR}) - \boldsymbol{Z}_i' \boldsymbol{b}$. For inference about $\boldsymbol{\beta}$, we describe an estimating equation approach based on linear rank statistics in the next subsection, and some other approaches are discussed in Section 6.4.3 following an illustrative example in Section 6.4.2.

6.4.1 Linear Rank Estimation of Regression Parameters

Linear rank statistics are usually defined as

$$S(\boldsymbol{b}) = \sum_{i=1}^{n} \boldsymbol{Z}_i \, c_i(\boldsymbol{b}) \tag{6.4}$$

and are often used for estimation of the regression parameter $\boldsymbol{\beta}$ in model (1.8) when right-censored failure time data are available (Jin et al., 2003; Kalbfleisch and Prentice, 2002). In (6.4), c_i is the score for the sample with $\boldsymbol{Z}_i$ that can take various forms and be either assigned or estimated. As discussed in Section 4.3.2, often, one chooses scores or estimates them so that $S(\boldsymbol{b})$ gives or approximates the score function of $\boldsymbol{\beta}$. Here we consider such an approach due to Betensky et al. (2001).

To choose scores $c_i(\boldsymbol{b})$ in (6.4), following the idea used in the previous section, we suppose that all Y_{ij}'s can be treated as arising from $K_1 + ... + K_n$ independent individuals. Then one has a set of current status data $\{Y_{ij} \, ; \, j = 1, ..., K_i \, , \, i = 1, ..., n\}$, which gives a likelihood function

$$L^*(\boldsymbol{\beta}, F) = \prod_{i=1}^{n} \prod_{j=1}^{K_i} \left[F(U_{ij}(\boldsymbol{\beta}))\right]^{Y_{ij}} \left[1 - F(U_{ij}(\boldsymbol{\beta}))\right]^{1-Y_{ij}} . \tag{6.5}$$

The resulting score function for $\boldsymbol{\beta}$ has the form

$$S^*(\boldsymbol{\beta}, F) = \sum_{i=1}^{n} \boldsymbol{Z}_i \sum_{j=1}^{K_i} \frac{-f(U_{ij}(\boldsymbol{\beta}))}{F(U_{ij}(\boldsymbol{\beta}))\,[1 - F(U_{ij}(\boldsymbol{\beta}))]} \left[Y_{ij} - F(U_{ij}(\boldsymbol{\beta}))\right] ,$$

where f denotes the derivative of F. This along with $E(Y_{ij} \,|\, \boldsymbol{Z}_i) = F(U_{ij}(\boldsymbol{\beta}))$ suggests that we can take $c_i(\boldsymbol{b}) = \sum_{j=1}^{K_i} \left[Y_{ij} - \hat{F}_{\boldsymbol{b}}(U_{ij}(\boldsymbol{b}))\right]$, which gives the linear rank statistic

$$S(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \left[Y_{ij} - \hat{F}_{\boldsymbol{b}}(U_{ij}(\boldsymbol{b}))\right] \boldsymbol{Z}_i \, . \tag{6.6}$$

In (6.6), $\hat{F}_{\boldsymbol{b}}$ denotes the NPMLE of F obtained from $L^*(\boldsymbol{\beta}, F)$ with fixed $\boldsymbol{\beta} = \boldsymbol{b}$, which can be determined by, for example, the pool adjacent violators algorithm as in Section 3.2.

It can be shown that under some regularity conditions, $\hat{F}_{\boldsymbol{b}}$ is a consistent estimate of F with $\boldsymbol{b} = \boldsymbol{\beta}$ and asymptotically $S(\boldsymbol{\beta})$ has expectation zero (Betensky et al., 2001). Thus it is natural to estimate $\boldsymbol{\beta}$ by a zero crossing of $S(\boldsymbol{b})$. For determination of the estimate, in practice, one can compute $S(\boldsymbol{b})$ for a fine grid of values of $\boldsymbol{b}$ and take the estimator $\hat{\boldsymbol{\beta}}_r$ as the value of $\boldsymbol{b}$ at which $S(\boldsymbol{b})$ is closest to zero. It should be noted that this approach may not be realistic if the dimension of $\boldsymbol{\beta}$ is high, and in this case, one may need to use some iterative algorithms.

For each index pair $(i\,j)$, let R_{ij} denote the set of index pairs $(k\,l)$ for which $\hat{F}_{\hat{\boldsymbol{\beta}}_r}(U_{kl}(\hat{\boldsymbol{\beta}}_r)) = \hat{F}_{\hat{\boldsymbol{\beta}}_r}(U_{ij}(\hat{\boldsymbol{\beta}}_r))$ and define

$$\hat{E}\left(\boldsymbol{Z}_i | U_{ij}(\hat{\boldsymbol{\beta}}_r)\right) = \frac{1}{|R_{ij}|} \sum_{(k\,l) \in R_{ij}} \boldsymbol{Z}_k$$

with $|R_{ij}|$ denoting the number of elements in R_{ij}. For estimation of the variance-covariance matrix of $S(\boldsymbol{\beta})$, for large n, Betensky et al. (2001) suggest

$$\hat{\Sigma}(\hat{\boldsymbol{\beta}}_r) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \sum_{j'=1}^{K_i} \left\{ \hat{F}_{\hat{\boldsymbol{\beta}}_r}(U_{ij}(\hat{\boldsymbol{\beta}}_r) \vee U_{ij'}(\hat{\boldsymbol{\beta}}_r)) \left[1 - \hat{F}(U_{ij}(\hat{\boldsymbol{\beta}}_r) \wedge U_{ij'}(\hat{\boldsymbol{\beta}}_r))\right] \right.$$

$$\left. \times \left[\boldsymbol{Z}_i - \hat{E}(\boldsymbol{Z}_i | U_{ij}(\hat{\boldsymbol{\beta}}_r))\right] \left[\boldsymbol{Z}_i - \hat{E}(\boldsymbol{Z}_i | U_{ij'}(\hat{\boldsymbol{\beta}}_r))\right]' \right\} .$$

Thus one can test $\boldsymbol{\beta} = 0$ using the statistic

$$X^2(\boldsymbol{\beta}) = S'(\boldsymbol{\beta})\, \hat{\Sigma}^{-1}(\boldsymbol{\beta})\, S(\boldsymbol{\beta})$$

with $\boldsymbol{\beta} = 0$ based on the χ^2 distribution with the degrees of freedom p and obtain an asymptotic $1 - \alpha$ level confidence set for $\boldsymbol{\beta}$ by

$$\{\, \boldsymbol{b};\, X^2(\boldsymbol{b}) < \chi^2_p(1-\alpha) \,\}$$

(Wei et al., 1990).

6.4.2 An Illustration

For purposes of comparison and illustration, we reanalyze the viral load data discussed in Section 6.3.2, which is data set III of Appendix A, using the estimation procedure described in the previous subsection. Let the T_i's and Z_i's be defined in as Section 6.3.2 and assume that the distribution of T_i can be described by the accelerated failure time model (1.8). To find the zero crossing of $S(b)$, as suggested in the previous subsection, we compute $S(b)$ for a grid of values of b. Figure 6.4 displays the curve of $S(b)$ given by the values obtained. This gives $\hat{\beta}_r = 1.6094$, similar to the estimates obtained in Section 6.3.2 using the proportional odds model.

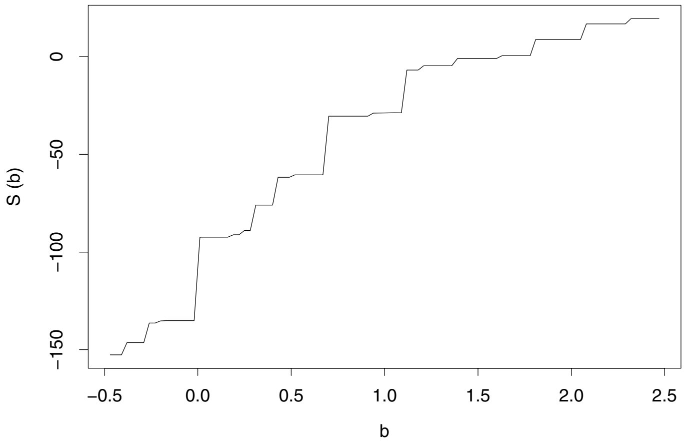

Fig. 6.4. Score function $S(b)$.

To test if there exists a difference between the two groups in terms of the time at which the RNA count drops below 500 viral copies/ml, we calculate the statistic $X^2(\beta)$ and obtain $X^2(\beta = 0) = 21.999$. This results in a p-value of almost zero for testing $\beta = 0$ based on the χ^2 distribution with degree of freedom 1. It suggests, as in Section 6.3.2, that the RNA counts reach below 500 copies much faster for the patients with initial RNA counts below 20000 than those with initial RNA counts above 20000. The asymptotic 95% confidence set given by the statistic $X^2(b)$ is $(0.6932, \infty)$, giving the same result.

6.4.3 Discussion

Rabinowitz et al. (1995) presented a class of different linear rank statistics based on the likelihood function $L(\boldsymbol{\beta}, F)$ given in (6.3) instead of the likelihood function $L^*(\boldsymbol{\beta}, F)$ given in (6.5). Specifically, their statistics have the form

$$S_R(\boldsymbol{b}) = \sum_{i=1}^{n} \frac{g[F(U_{iR}(\boldsymbol{b}))] - g[F(U_{iL}(\boldsymbol{b}))]}{F(U_{iR}(\boldsymbol{b})) - F(U_{iL}(\boldsymbol{b}))} \, \boldsymbol{Z}_i \tag{6.7}$$

with F replaced by the maximum likelihood estimate of F derived from $L(\boldsymbol{\beta}, F)$. In these statistics, g is a known weight function that can be chosen to minimize the asymptotic variance of linear functions of regression parameter estimates. Note that if we let $g = f \circ F^{-1}$, the statistic S_R then becomes the score statistic of $\boldsymbol{\beta}$ given by $L(\boldsymbol{\beta}, F)$. A major advantage of the statistics given in (6.7) is that with the optimal weight function and suitable regularity conditions, asymptotic efficiency can be achieved. However, the numerical and computational effort involved can be too much for the approach to be practical (Betensky et al., 2001).

For the rank statistics defined in both (6.6) and (6.7), there could exist more than one zero crossing. For this, Rabinowitz et al. (1995) suggested using the zero crossing that is in a neighborhood of an $n^{-1/2}$-consistent estimator of $\boldsymbol{\beta}$. One way to obtain such an estimator is to consider the current status data given by only the first observation times U_{i1}'s and to maximize the resulting likelihood function

$$\prod_{i=1}^{n} \left[F(U_{i1}(\boldsymbol{\beta}))\right]^{Y_{i1}} \left[1 - F(U_{i1}(\boldsymbol{\beta}))\right]^{1-Y_{i1}}$$

over $\boldsymbol{\beta}$ and F together.

To derive a rank statistic, one can also directly rank the censored intervals similar to ranking the survival times, T_i, when they are known (Li, 2003; Li and Pu, 1999). For example, Li and Pu (2003) investigated such an approach and proposed the statistic

$$S_{LP}(b) = \sum_{i<j} \left[I(Z_i < Z_j) - I(Z_i > Z_j) \right] \left[I(U_{iR}(b) < U_{jL}(b)) \right.$$
$$\left. - I(U_{iL}(b) > U_{jR}(b)) \right] .$$

In this statistic, it is assumed that there exists only one covariate. That is, the Z_i's and β are scalars. Note that $S_{LP}(b)$ is discrete and has either no zero crossing or multiple zero crossings. Define $\hat{\beta}_1 = \sup\{ b : S_{LP}(b) \geq 0 \}$ and $\hat{\beta}_2 = \inf\{ b : S_{LP}(b) \leq 0 \}$. They suggested estimating β by $\hat{\beta}_{LP} = (\hat{\beta}_1 + \hat{\beta}_2)/2$ and showed that $n^{1/2}(\hat{\beta}_{LP} - \beta)$ has an asymptotic normal distribution with mean zero. In contrast, no similar rigorous investigation of the estimates defined using rank statistics (6.6) or (6.7) is available yet. A shortcoming of S_{LP} is that it only applies to single covariate situations and it is not straightforward to generalize it to multivariate covariate situations. Also it is not hard to see that the statistic S_{LP} could be very inefficient because many terms in the summation could be zero if, for example, the data contain no or few exact observations.

As commented before, one can always apply the maximum likelihood approach for fitting the accelerated failure time model to general interval-censored data. However, unlike the PH model or the proportional odds model, use of the maximum likelihood approach here is relatively hard. The main difficulty is that the regression parameter $\boldsymbol{\beta}$ and the nuisance function F are tangled in the likelihood function. Consequently, the profile likelihood function is not a smooth function of $\boldsymbol{\beta}$ and also the log likelihood function is not twice-differentiable with respect to $\boldsymbol{\beta}$. Huang and Wellner (1997) investigated this and proved the consistency of the maximum likelihood estimators of $\boldsymbol{\beta}$ and F. However, no asymptotic distribution theory for the estimators is available yet, even for the case of current status data.

6.5 Analysis with the Logistic Model

In this section, we discuss the analysis of discrete interval-censored failure time data and consider fitting the logistic model (1.12) to the data. As in Section 5.5, let $0 < s_1 < ... < s_{m+1}$ denote all possible values that the survival times T_i's of interest take and suppose that the observed data are given in the form (6.1). As in Section 6.2, define $\alpha_{ij} = I(s_j \in (L_i, R_i])$, $j = 1, ..., m+1$, $i = 1, ..., n$. Then we have the likelihood function proportional to

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \sum_{j=1}^{m+1} \alpha_{ij} \left[\prod_{k=0}^{j-1} \left(1 + \gamma_k e^{\boldsymbol{Z}_i' \boldsymbol{\beta}}\right)^{-1} - \prod_{k=0}^{j} \left(1 + \gamma_k e^{\boldsymbol{Z}_i' \boldsymbol{\beta}}\right)^{-1} \right],$$

where $\gamma_0 = 0$, $\gamma_{m+1} = \infty$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_m)'$. As before, for each j, define $\alpha_j = \log \gamma_j$. The log likelihood function then has the form

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{m+1} \alpha_{ij} \left(\prod_{k=0}^{j-1} \frac{1}{1 + e^{\alpha_k + \boldsymbol{Z}_i' \boldsymbol{\beta}}} - \prod_{k=0}^{j} \frac{1}{1 + e^{\alpha_k + \boldsymbol{Z}_i' \boldsymbol{\beta}}} \right) \right]$$

in terms of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, where $\alpha_0 = -\infty$, $\alpha_{m+1} = \infty$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$.

For inference, we first discuss the maximum likelihood approach in the next subsection. Score tests for survival comparison are then considered with the focus on their relationships to some other tests. Finally, two illustrative examples are provided.

6.5.1 Maximum Likelihood Estimation of Parameters

This subsection considers estimation of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ using the maximum likelihood approach. For this, we have

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{Z}_i \, g_i^{-1} \, h_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

and

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} = \sum_{i=1}^{n} g_i^{-1} \, p_j(\boldsymbol{Z}_i) \, c_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}),$$

$j = 1, ..., m$. In the expressions above, $h_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{j=1}^{m+1} \alpha_{ij} \, h_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})$,

$$h_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = S(s_j; \boldsymbol{Z}_i) \sum_{k=0}^{j} p_k(\boldsymbol{Z}_i) - S(s_{j-1}; \boldsymbol{Z}_i) \sum_{k=0}^{j-1} p_k(\boldsymbol{Z}_i),$$

$$p_j(\boldsymbol{Z}_i) = \frac{e^{\alpha_j + \boldsymbol{Z}_i' \boldsymbol{\beta}}}{1 + e^{\alpha_j + \boldsymbol{Z}_i' \boldsymbol{\beta}}}, \; p_0(\boldsymbol{Z}_i) = 0, \; p_{m+1}(\boldsymbol{Z}_i) = 1,$$

$$c_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha}) \;=\; \alpha_{ij}\, S(s_j;\boldsymbol{Z}_i) \;-\; \sum_{k=j+1}^{m+1} \alpha_{ik}\, f_{ik}(\boldsymbol{\beta},\boldsymbol{\alpha})\,,$$

$$f_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha}) \;=\; S(s_{j-1};\boldsymbol{Z}_i) \;-\; S(s_j;\boldsymbol{Z}_i)\,,$$

$S(s_j;\boldsymbol{Z}_i)$ is defined in model (1.12), and $g_i \;=\; \sum_{j=1}^{m+1} \alpha_{ij}\, f_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha})$ as in Section 6.2.1. Thus it is natural to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ defined as the solution to score equations

$$\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \;=\; 0 \;\; \text{and} \;\; \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \alpha_j} \;=\; 0\,,$$

which can be solved by the Newton-Raphson algorithm.

The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, as in Section 5.5.1, can be estimated by the inverse of the observed Fisher information matrix

$$I(\boldsymbol{\beta},\boldsymbol{\alpha}) \;=\; \begin{pmatrix} I_{11} \; I_{12} \\ I_{21} \; I_{22} \end{pmatrix}$$

with $\boldsymbol{\beta} \;=\; \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha} \;=\; \hat{\boldsymbol{\alpha}}$. Here we have that

$$I_{11} = -\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i' \left\{ g_i^{-1} \sum_{j=1}^{m+1} \alpha_{ij} \left[S(s_j;\boldsymbol{Z}_i) \left(\sum_{k=0}^{j} p_k(\boldsymbol{Z}_i) \right)^2 \right.\right.$$
$$-S(s_j;\boldsymbol{Z}_i) \sum_{k=0}^{j} p_k(\boldsymbol{Z}_i)\,(1-p_k(\boldsymbol{Z}_i)) - S(s_{j-1};\boldsymbol{Z}_i) \left(\sum_{k=0}^{j-1} p_k(\boldsymbol{Z}_i) \right)^2$$
$$\left.\left. +S(s_{j-1};\boldsymbol{Z}_i) \sum_{k=0}^{j-1} p_k(\boldsymbol{Z}_i)\,(1-p_k(\boldsymbol{Z}_i)) \right] + \left[\frac{h_i(\boldsymbol{\beta},\boldsymbol{\alpha})}{g_i} \right]^2 \right\}$$

and $I_{12} \;=\; I_{21}'$ and I_{22} have elements

$$-\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \alpha_j \partial \boldsymbol{\beta}} \;=\; \sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ g_i^{-1} p_j(\boldsymbol{Z}_i) \left[\alpha_{ij} S(s_j;\boldsymbol{Z}_i) \sum_{k=0}^{j} p_k(\boldsymbol{Z}_i) \right.\right.$$
$$\left.\left. -c_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha})(1-p_j(\boldsymbol{Z}_i)) + \sum_{k=j+1}^{m+1} \alpha_{ik} h_{ik}(\boldsymbol{\beta},\boldsymbol{\alpha}) \right] + \frac{h_i(\boldsymbol{\beta},\boldsymbol{\alpha}) c_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha}) p_j(\boldsymbol{Z}_i)}{g_i^2} \right\}\,,$$

$$-\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \alpha_j^2} \;=\; \sum_{i=1}^{n} \left\{ \frac{p_j(\boldsymbol{Z}_i)(2p_j(\boldsymbol{Z}_i)-1) c_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha})}{g_i} + \left[\frac{p_j(\boldsymbol{Z}_i) c_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha})}{g_i} \right]^2 \right\}\,,$$

and

$$-\frac{\partial^2 l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_k} = \sum_{i=1}^{n} \frac{p_j(\boldsymbol{Z}_i) p_k(\boldsymbol{Z}_i) c_{ik}(\boldsymbol{\beta},\boldsymbol{\alpha})}{g_i^2} \left[\alpha_{ij} S(s_j;\boldsymbol{Z}_i) + \sum_{l=1}^{j} \alpha_{il} f_{il}(\boldsymbol{\beta},\boldsymbol{\alpha}) \right]$$

for $j \;<\; k$, respectively, $j\,,\; k \;=\; 1,...,m$.

6.5.2 Score Tests for Comparison of Survival Functions

In this subsection, we suppose that study subjects come from $p+1$ different groups with the $\boldsymbol{Z}_i$'s being the group indicators, and the goal is to test whether the subjects in different groups share the same survival function. For this, as discussed before, one can apply the score test with the score test statistic given by $U_L = \partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}$ with $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_0$, the maximum likelihood estimator of $\boldsymbol{\alpha}$ with $\boldsymbol{\beta} = 0$. As for U_{PH} in Section 6.2, the variance-covariance matrix of U_L can be estimated by $I_{11|2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$ with $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_0$ when n is large (Sun, 1997a).

To see the relationship between U_L and some other test statistics described in the previous sections, using algebra, it can be shown that U_L can be rewritten as

$$U_L = \sum_{i=1}^{n} \boldsymbol{Z}_i \sum_{j=1}^{m+1} \left(\frac{\alpha_{ij} \hat{f}_j}{\hat{g}_i} - \frac{\hat{p}_j}{\hat{g}_i} \sum_{l=j}^{m+1} \alpha_{il} \hat{f}_l \right) .$$

In this expression, $\hat{f}_j$, $\hat{g}_i$, and $\hat{p}_j$ denote $f_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})$, g_i, and $p_j(\boldsymbol{Z}_i)$ defined above with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ set to be equal to 0 and $\hat{\boldsymbol{\alpha}}_0$, respectively. For each j, define $D_j = \{ i \, ; \, \alpha_{ij} = 1 \}$ and $R_j = \{ i \, ; \, \alpha_{ij'} \geq 1 \text{ for some } j' \geq j \}$, the index sets of subjects who may fail at s_j and at or after s_j given the observed data, respectively. Also define

$$d_{ij} = \frac{\alpha_{ij} \, \hat{f}_j}{\hat{g}_i} \ , \ r_{ij} = \sum_{l=j}^{m+1} \frac{\alpha_{il} \, \hat{f}_l}{\hat{g}_i} .$$

They represent the estimated probabilities that the survival event of interest from subject i occurs at s_j and at or after s_j given the observed data and that there is no survival difference among the $p+1$ groups. Then U_L has the form

$$U_L = \sum_{j=1}^{m} \left(\sum_{i \in D_j} \boldsymbol{Z}_i \, d_{ij} - \hat{p}_j \sum_{i \in R_j} \boldsymbol{Z}_i \, r_{ij} \right) \tag{6.8}$$

and it can be shown that $\hat{p}_j = d_j / r_j = \sum_{i \in D_j} d_{ij} / \sum_{i \in R_j} r_{ij}$, $j = 1, ..., m$.

For the two-sample comparison problem with $Z_i = 0$ or 1, the statistic U_L reduces to $U_L = \sum_{j=1}^{m} (d_{1j} - \hat{p}_j \, r_{1j})$, where

$$d_{1j} = \sum_{\{i \, : \, i \in D_j, Z_i = 1\}} d_{ij} \text{ and } r_{1j} = \sum_{\{i \, : \, i \in R_j, Z_i = 1\}} r_{ij} .$$

The quantities d_{1j} and r_{1j} can be regarded as estimates of the numbers of failures and risks, respectively, at s_j for subjects in the group with $Z_i = 1$. Thus, U_L is a summation of the estimated numbers of failures assuming the two groups may have different failure distributions minus the same estimated numbers of failures, but under the no difference assumption, among subjects in the group with $Z_i = 1$.

It can be seen from the expression (6.8) that U_L is similar to the generalized log-rank test statistic U_r given in Section 4.3 and the two have similar meanings. However, we remark that the two test statistics are different unless there are no right-censored observations, which implies that U_L does not reduce to the log-rank test statistic in the case of right-censored data. This is because in U_L, right-censored subjects are treated to be at risk at all times, while they are regarded to be at risk only up to the right censoring times in U_r and the log-rank test statistic.

Using the notation defined above, we can rewrite the score test statistic U_{PH} given in Section 6.2 as

$$U_{PH} = \sum_{i=1}^{n} \boldsymbol{Z}_i \sum_{j=1}^{m+1} \left[\frac{\log(1-\hat{p}_j)}{\hat{g}_i} \sum_{l=j}^{m+1} \alpha_{il} \hat{f}_l - \frac{\alpha_{ij} \hat{f}_j \log(1-\hat{p}_j)}{\hat{g}_i \, \hat{p}_j} \right]$$

$$= \sum_{j=1}^{m} \frac{\log(1-\hat{p}_j)^{-1}}{\hat{p}_j} \left(\sum_{i \in D_j} \boldsymbol{Z}_i \, d_{ij} - \hat{p}_j \sum_{i \in R_j} \boldsymbol{Z}_i \, r_{ij} \right) . \qquad (6.9)$$

It is apparent from (6.8) and (6.9) that U_L and U_{PH} are different. However, they would be close to each other if the d_j's, the estimated numbers of failures, are small relative to the r_j's, the estimated numbers of risks. In this case, $\hat{p}_j$ is close to zero and thus $-\log(1-\hat{p}_j)$ is close to $\hat{p}_j$. As U_L, U_{PH} also does not agree with the log-rank test statistic in the case of right-censored data. Sun (1997a) studied a more general logistic model that results in a class of weighted score tests that have the form of U_{PH} with $\log(1-\hat{p}_j)^{-1}/\hat{p}_j$ replaced by a weight function.

6.5.3 Two Examples

In this subsection, we apply the inference procedures presented in the previous subsections to two sets of discrete interval-censored failure time data. First we analyze the grouped breast cancer data given by grouping the breast cancer data presented in Table 1.4 and assuming that the breast retraction can occur only at 6, 12, 18, 24, 30, 36, 42, 48, 54, and 65 months, respectively. We then discuss the data given in Table 5.4 about tumor occurrence rates.

For the grouped breast cancer data, let the T_i's and Z_i's be defined as in Section 6.2.3 and assume that T_i can be described by the logistic model (1.12). In this case, we have $m = 9$. The maximum likelihood estimation procedure discussed in Section 6.5.1 yields $\hat{\beta} = 1.0128$ with estimated standard error equal to 0.3307 and a resulting p-value of 0.002 for testing $\beta = 0$. For the score test, we obtain $U_L = 9.7826$ and its estimated standard deviation is 3.1144, also giving a p-value of 0.002 for the comparison of the two treatment groups. The results are similar to those given in Section 6.2.3 using the PH model and indicate that the patients in the RCT group had significantly higher risk to develop breast retraction than those in the RT group.

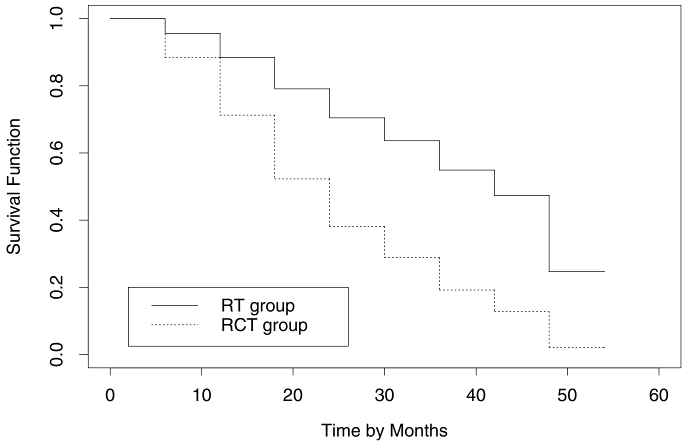

Fig. 6.5. Estimated survival functions for breast cancer data.

Figure 6.5 presents the estimated survival functions of the time to breast retraction for the patients in the two groups under model (1.12). They are similar to the corresponding estimates given in Figure 6.1 and suggest that as with the PH model, the logistic model also seems to be reasonable for the data. To investigate the effect of grouping on the analysis, we repeated the analysis above by assuming that the breast retraction can occur only at 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, and 65 months, respectively, giving $m = 12$. The estimates obtained are $\hat{\beta} = 0.9844$ and $U_L = 10.3030$ with their estimated standard errors being 0.3195 and 3.2536, respectively. They are similar to those given above.

For comparison, we consider the tumor occurrence data discussed in Section 5.5.2. For each animal, as in Section 5.5.2, define T_i to be the tumor occurrence time and assume that the tumor occurs only at the end of each interval. Also define $Z_i = 0$ for male rats and 1 for female rats. The inference procedure described in the previous subsections gives $\hat{\beta} = 0.9922$ and $U_L = 12.4478$ with their estimated standard deviations being 0.3636 and 3.6067, respectively. Comparing tumor occurrence rates between male and female rats, we obtain p-values of 0.007 and 0.0006, respectively, which indicate that the female rats had significantly higher tumor occurrence rate than male rats. These are similar to the results given in Section 5.5.2 except that the score test based on U_L suggests a more significant difference than others. Figure 6.6 displays estimates of the survival functions of the time to tumor occurrence obtained under the logistic model for both male and female rats. They are similar to those presented in Figure 5.5 that were obtained using the grouped PH model.

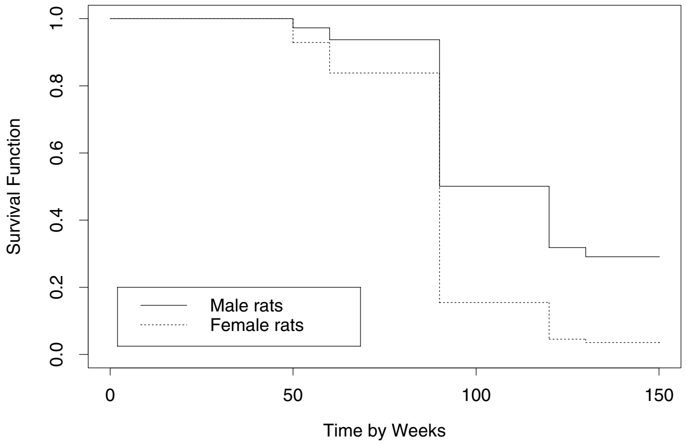

Fig. 6.6. Estimated survival functions for tumor data.

6.6 Bibliography, Discussion, and Remarks

The literature on regression analysis of case II interval-censored failure time data goes back a long way with most of early work on grouped data (Pierce et al., 1979; Prentice and Gloeckler, 1978). It wasn't until the mid-1980s that many articles about general interval-censored data began to appear, including the seminal article Finkelstein (1986), which is the first that studied the use of the PH model for interval-censored data. Others that investigated the PH model include Alioum and Commenges (1996), Betensky et al. (2002), Cai and Betensky (2003), Datta et al. (2000), Goggins et al. (1998), Huang and Wellner (1997), Huber-Carol and Vonta (2004), Kim (2003a, b), Pan and Chappell (2002), and Satten (1996). In particular, Huang and Wellner (1997) discussed the asymptotic properties of the PH model along with other models. Alioum and Commenges (1996), Datta et al. (2000), Huber-Carol and Vonta (2004), and Pan and Chappell (2002) studied the use of the PH model for the analysis of failure time data that involve interval censoring as well as truncation.

Several other semiparametric models have also been considered for regression analysis of case II interval-censored data. These include the proportional odds model, which was studied by Huang and Rossini (1997), Huang and Wellner (1997), Rabinowitz et al. (2000), and Shen (1998), and the accelerated failure time model, for which references include Betensky et al. (2001), Li and Pu (1999, 2003), Rabinowitz et al. (1995), and Xue et al. (2006). Also for the problem, Kooperberg and Clarkson (1997) discussed linear spline models, and Sun (1997a) investigated the logistic model. Furthermore, Younes and

Lachin (1997) and Zhang et al. (2005) proposed to use the linear transformation model (1.9), and Zeng et al. (2006) recently considered the additive hazards model.

Others that discussed the topic include Bacchetti and Quale (2002), Carstensen (1996), Finkelstein and Wolfe (1985, 1986), Joly et al. (1998), and Yu and Wong (2003). Among them, Bacchetti and Quale (2002) and Joly et al. (1998) considered regression analysis of interval-censored and truncated data using the penalized likelihood approach under some general regression models. All the regression approaches discussed in this and other chapters model the survival time of interest conditional on covariates. To model the survival time marginally, Finkelstein and Wolfe (1985, 1986) proposed to use a parametric model for the conditional distribution of covariates given the survival variable.

In both this chapter and Chapter 5, it has been suggested to use the observed Fisher information matrix to estimate variance-covariance matrix of the maximum likelihood estimators of regression parameters. As an alternative, one can also apply the profile likelihood approach, which estimates the variance-covariance matrix by the inverse of the curvature of the profile likelihood (Huang and Wellner, 1997). This method is feasible if the dimension of the regression parameters is low and requires that the profile likelihood is a smooth function of the regression parameters. As with the observed Fisher information approach, the asymptotic validity of the profile likelihood approach also needs to be verified for each situation.

For estimation of the finite-dimensional regression parameter in a semiparametric model, in addition to the approaches discussed in the previous sections, another commonly used approach is to base the estimation on the efficient score function of the parameter (Bickel et al., 1993; van der Vaart, 1998). Among others, Huang and Wellner (1977) applied it to both case I and II interval-censored data under several semiparametric regression models discussed in this chapter and Chapter 5. Martinussen and Scheike (2002b) considered this approach for regression analysis of current status data using the additive hazards model. A key advantage of this efficient score approach is that the resulting estimators are efficient. The efficient score function is also needed for calculation of the information bound for the parameter. On the other hand, the derivation of an efficient score function is usually not an easy task, and use of it requires estimating an infinite-dimensional nuisance parameter such as the cumulative hazard function in the PH model. This latter feature makes the approach unattractive compared with the approaches that do not need estimation of the nuisance parameter.

For any inference problem about interval-censored failure time data, a natural approach is to generalize the existing approaches developed for the same problem with respect to right-censored data, and many approaches discussed in this book were indeed developed this way. Of course, such methods are usually more complicated in terms of both implementation and investigation of their properties than their counterparts for right-censored data. On the other hand, there exist many inference approaches for right-censored data that seem impossible to generalize, or are still not available, for interval-censored data due to the complexity of the data and censoring structure involved. One such example is the partial likelihood approach (Cox, 1972; Kalbfleisch and Prentice, 2002), which provides a simple and efficient method for regression analysis of right-censored data using the PH model. It also allows one to use the counting process and associated martingale theory to easily establish theoretical justification of the approach (Andersen et al., 1993; Andersen and Gill, 1982). Unfortunately, the partial likelihood approach does not seem to be directly applicable to interval-censored data, and no method similar is available yet.

There exist many challenging unsolved inference problems about interval-censored data. Of primary importance is the investigation of the asymptotic properties of the maximum likelihood estimator for various semiparametric regression models including those discussed in this chapter. For the PH model and the proportional odds model, the consistency, asymptotic efficiency, and asymptotic normality of maximum likelihood estimators of regression parameters have been established (Huang and Wellner, 1997). But for other models like the accelerated failure time model and the additive hazards model, these properties are still unknown. For these models, inferences about the regression parameter are based on approximate likelihood or estimating equations. For these approximate likelihood or estimating equation approaches, one needs to investigate their efficiency, which is usually more challenging than developing the approximate likelihood or estimating equations. Of course, as remarked above, it would be really useful if a method similar to the partial likelihood approach were available.

For case II interval-censored data problems, a general approach to develop estimating equations or inference procedures is to transfer them to current status data problems as in Section 6.4.1. Of course, this assumes that a simple inference procedure is available in the case of current status data and the data are recorded in representations (1.2) or (1.3). The same idea was also used in Section 4.4.3 for developing test statistics for survival comparison. For example, for case II interval-censored data given in representation (1.2), an estimating equation or inference procedure can be easily obtained as a linear combination of those based on current status data on U and V, respectively. Of course, the asymptotic validity and properties of the resulting methods need to be investigated as well as their efficiency, which is often a difficult task and for which little research is available.

As discussed in Section 5.6, time-dependent covariates often arise in practice. However, most of the inference approaches developed for interval-censored data only apply to time-independent covariates. Some exceptions include van der Laan and Robins (1998), Lin, Oakes, and Ying (1998), and Martinussen and Scheike (2002b). To compare several survival functions under semiparametric models, in addition to using score tests, another natural approach is to apply the likelihood ratio type test. As with the asymptotic distribution of the maximum likelihood estimator, however, the derivation of the asymptotic distribution of the likelihood ratio test statistic is also not an easy task. As discussed in Chapter 5, model selection is another important and difficult topic for regression analysis of interval-censored data.

7

Analysis of Bivariate Interval-censored Data

7.1 Introduction

Bivariate failure time data occur in many situations. A standard example is studies on twins or eyes where one is interested in times to the occurrences in both twins or eyes of a certain event such as some diseases or disease-related symptoms. By bivariate failure time data, we usually mean that there exist two failure time variables of interest, and the two variables cannot be assumed to be independent. For example, in an eye study, the two variables could be times to the blindness for both left and right eyes, and the two are obviously related. A more general example of bivariate failure time data is times to two same or different types of events that happen on the same subject. It is apparent that bivariate failure time data are special cases of multivariate failure time data that concern information about several possibly related failure times. Sometimes, multivariate failure time data are also referred to as correlated failure time data. This chapter focuses on bivariate failure time data in the presence of interval censoring.

A number of authors have studied the analysis of bivariate failure time data when only right censoring is present. These include Cai and Kim (2003), Cai and Prentice (1995), Li and Lagakos (2004), Lin (1994), Lin and Ying (1993), Prentice and Cai (1992), Prentice and Kalbfleisch (2003), and Wei et al. (1989). In particular, Hougaard (2000) is an excellent book about the analysis of multivariate failure time data. As discussed before, the mechanism that generates interval censoring is usually much more complicated and difficult to deal with than that behind right censoring. For the analysis of bivariate interval-censored data, one has to deal with the difficulties that exist for the analysis of univariate interval-censored data as well as those caused by the correlation structure between two related failure times.

Several inference problems for bivariate failure time data are discussed in this chapter. In Section 7.2, we first discuss estimation of the association parameter of two related failure variables in the presence of interval censoring. For this problem, we focus attention on the situation where the joint survival function of the two failure variables can be described by a copula model. A two-stage estimation procedure given in Sun, Wang, and Sun (2006) and Wang and Ding (2000) is described. Section 7.3 deals with maximum likelihood estimation of a joint distribution function for two failure times of interest. As in the case of univariate interval-censored data, determination of the maximum likelihood estimator consists of two steps: determination of the regions where probability masses lie and maximization of the likelihood. Although there is basically no difference between the second steps in the two situations, the first step is much more complicated for bivariate interval-censored data. Three algorithms for the first step are discussed. Regression analysis under the discrete or grouped PH model is the topic for Section 7.4, and the marginal inference approach commonly used for multivariate right-censored failure time data is discussed. Section 7.5 gives bibliographic notes and general discussion and remarks about the analysis of bivariate interval-censored data, in particular some issues and inference approaches not discussed in the previous sections. As in the previous chapters, the censoring mechanism is assumed to be independent of the underlying failure variables in this chapter.

7.2 Estimation of the Association Parameter

Let T_1 and T_2 denote two possibly correlated failure times. In the analysis of bivariate failure time data, one of the main interests is measuring the dependence or association of T_1 and T_2. Several approaches can be used to characterize this association, and for right-censored failure time data, a few methods have been proposed for making inferences about the association. In this section, we focus attention on situations for which the joint survival function of T_1 and T_2 can be described by a copula model, a commonly used model for bivariate failure time data (Clayton, 1978; Fine and Jiang, 2000; Genest and Rivest, 1993; Hougaard, 1986).

7.2.1 The Copula Model and the Likelihood Function

Let $S_1(t)$ and $S_2(t)$ denote the marginal survival functions of T_1 and T_2, respectively, and $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ their joint survival function. A copula model assumes that $S(t_1, t_2)$ can be expressed as

$$S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2)) \,, \tag{7.1}$$

where C_α is a distribution function on the unit square, and $\alpha \in R$ is a global association parameter.

One attractive feature of model (7.1) is flexibility. In fact, it includes as special cases many useful bivariate failure time models such as the Archimedean copula family,

$$C_\alpha(u, v) = \phi_\alpha\{\phi_\alpha^{-1}(u) + \phi_\alpha^{-1}(v)\} \,,\; 0 \leq u \,,\, v \leq 1 \,,$$

where $0 \le \phi_\alpha \le 1$, $\phi_\alpha(0) = 1$, $\phi'_\alpha < 0$, $\phi''_\alpha > 0$. In particular, taking $\phi_\alpha(u) = (1+u)^{1/(1-\alpha)}$, the Laplace transformation of a gamma distribution, one has

$$C_\alpha(u,v) = (u^{1-\alpha} + v^{1-\alpha} - 1)^{1/(1-\alpha)}\,,\ \alpha > 1\,,$$

which is commonly referred to as the Clayton family (Clayton, 1978). Suppose that $S(t_1,t_2)$ is absolutely continuous. Then under the Clayton model, α also represents the ratio of the hazard function of $T_1 = t_1$ given $T_2 = t_2$ to that given $T_2 \ge t_2$, or that of $T_2 = t_2$ given $T_1 = t_1$ against given $T_1 \ge t_1$.

Another attractive feature of copula models is that the marginal distributions do not depend on the choice of the association structure. Thus one can model the marginal distributions and the association separately.

A second common way to characterize the association of two random variables is to use Kendall's τ defined as

$$\tau = Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\} - Pr\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\}$$

for i.i.d. replicates (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) of (T_1, T_2). It represents the difference between the probabilities of concordance and discordance and Betensky and Finkelstein (1999a) considered this approach. Under model (7.1), Kendall's τ can be expressed as

$$\tau = 4\int_0^1\int_0^1 C_\alpha(u,v)\,C_\alpha(du\,,\,dv) - 1\,, \tag{7.2}$$

and thus for a fixed copula model, C_α, one only needs to consider estimation of α.

To describe the observed data, suppose that there exist two pairs of random variables $(U^{(1)}, V^{(1)})$ and $(U^{(2)}, V^{(2)})$, representing monitoring times for T_1 and T_2, respectively. Define

$$\Delta_1^{(j)} = I(T_j \le U^{(j)})\ \ ,\ \ \Delta_2^{(j)} = I(U^{(j)} < T_j \le V^{(j)})\,,$$

$j = 1, 2$. The independent censoring mechanism implies that (T_1, T_2) is independent of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$, but $(U^{(1)}, V^{(1)})$ and $(U^{(2)}, V^{(2)})$ could be dependent. Let $H(\boldsymbol{x})$ denote the joint distribution function of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$ and $G_\alpha(\boldsymbol{x}, \boldsymbol{\delta})$ the subdistribution function of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \boldsymbol{\Delta})$, where $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$, $\boldsymbol{\delta} = (\delta_1^{(1)}, \delta_2^{(1)}, \delta_1^{(2)}, \delta_2^{(2)})$ and $\boldsymbol{\Delta} = (\Delta_1^{(1)}, \Delta_2^{(1)}, \Delta_1^{(2)}, \Delta_2^{(2)})$. The density or probability functions of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$ and $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \boldsymbol{\Delta})$ will be denoted by $h(\boldsymbol{x})$ and $g_\alpha(\boldsymbol{x}, \boldsymbol{\delta})$.

Suppose that one observes n i.i.d. replicates of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \boldsymbol{\Delta})$ given by

$$\{\,U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \boldsymbol{\Delta}_i\ ;\ i = 1, ..., n\,\}\,,$$

where $\boldsymbol{\Delta}_i = (\Delta_{1i}^{(1)}, \Delta_{2i}^{(1)}, \Delta_{1i}^{(2)}, \Delta_{2i}^{(2)})$. We have bivariate current status data if $U_i^{(1)} = V_i^{(1)}$, $U_i^{(2)} = V_i^{(2)}$, and $\Delta_{2i}^{(1)} = \Delta_{2i}^{(2)} = 0$. Define

$$S_{11}(\alpha,\boldsymbol{x}) = P(T_1 \le x_1, T_2 \le x_3) = 1 - S_1(x_1) - S_2(x_3) + C_\alpha(S_1(x_1), S_2(x_3))\,,$$

$$S_{12}(\alpha,\boldsymbol{x}) = P(T_1 \le x_1, x_3 < T_2 \le x_4) = S_2(x_3) - S_2(x_4) + C_\alpha(S_1(x_1), S_2(x_4)) \\ -C_\alpha(S_1(x_1), S_2(x_3))\,,$$

$$S_{13}(\alpha,\boldsymbol{x}) = P(T_1 \le x_1, T_2 > x_4) = S_2(x_4) - C_\alpha(S_1(x_1), S_2(x_4))\,,$$

$$S_{21}(\alpha,\boldsymbol{x}) = P(x_1 < T_1 \le x_2, T_2 \le x_3) = S_1(x_1) - S_1(x_2) + C_\alpha(S_1(x_2), S_2(x_3)) \\ -C_\alpha(S_1(x_1), S_2(x_3))\,,$$

$$S_{22}(\alpha,\boldsymbol{x}) = P(x_1 < T_1 \le x_2, x_3 < T_2 \le x_4) = C_\alpha(S_1(x_1), S_2(x_3)) \\ -C_\alpha(S_1(x_1), S_2(x_4)) - C_\alpha(S_1(x_2), S_2(x_3)) + C_\alpha(S_1(x_2), S_2(x_4))\,,$$

$$S_{23}(\alpha,\boldsymbol{x}) = P(x_1 < T_1 \le x_2, T_2 > x_4) = C_\alpha(S_1(x_1), S_2(x_4)) \\ -C_\alpha(S_1(x_2), S_2(x_4))\,,$$

$$S_{31}(\alpha,\boldsymbol{x}) = P(T_1 > x_2, T_2 \le x_3) = S_1(x_2) - C_\alpha(S_1(x_2), S_2(x_3))\,,$$

$$S_{32}(\alpha,\boldsymbol{x}) = P(T_1 > x_2, x_3 < T_2 \le x_4) = C_\alpha(S_1(x_2), S_2(x_3)) \\ -C_\alpha(S_1(x_2), S_2(x_4))$$

and

$$S_{33}(\alpha,\boldsymbol{x}) = P(T_1 > x_2, T_2 > x_4) = C_\alpha(S_1(x_2), S_2(x_4))\,.$$

Then under model (7.1), the log likelihood function is given by

$$l(\alpha, S_1, S_2) = \sum_{i=1}^{n} l(\alpha, S_1, S_2, U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \boldsymbol{\Delta}_i)\,,$$

where

$$l(\alpha, S_1, S_2, \boldsymbol{x}, \boldsymbol{\delta}) = \delta_1^{(1)}\delta_1^{(2)} \log S_{11}(\alpha,\boldsymbol{x}) + \delta_1^{(1)}\delta_2^{(2)} \log S_{12}(\alpha,\boldsymbol{x}) \\ +\delta_1^{(1)}(1 - \delta_1^{(2)} - \delta_2^{(2)}) \log S_{13}(\alpha,\boldsymbol{x}) + \delta_2^{(1)}\delta_1^{(2)} \log S_{21}(\alpha,\boldsymbol{x}) \\ +\delta_2^{(1)}\delta_2^{(2)} \log S_{22}(\alpha,\boldsymbol{x}) + \delta_2^{(1)}(1 - \delta_1^{(2)} - \delta_2^{(2)}) \log S_{23}(\alpha,\boldsymbol{x}) \\ +(1 - \delta_1^{(1)} - \delta_2^{(1)})\delta_1^{(2)} \log S_{31}(\alpha,\boldsymbol{x}) + (1 - \delta_1^{(1)} - \delta_2^{(1)})\delta_2^{(2)} \log S_{32}(\alpha,\boldsymbol{x}) \\ +(1 - \delta_1^{(1)} - \delta_2^{(1)})(1 - \delta_1^{(2)} - \delta_2^{(2)}) \log S_{33}(\alpha,\boldsymbol{x}) + \log h(\boldsymbol{x})\,.$$

In the next subsection, we discuss estimation of the association parameter α.

7.2.2 A Two-Stage Estimation Procedure

A natural estimator of α is obtained by maximizing the log likelihood function, $l(\alpha, S_1, S_2)$, but involves estimating α, S_1, and S_2 together. On the other hand, it is much easier to estimate the marginal survival functions S_1 and S_2 separately using the univariate interval-censored data on T_1 and T_2, respectively. This motivates the following two-stage estimation procedure: first estimate S_1 and S_2 separately and then estimate α by maximizing the pseudo log likelihood given by $l(\alpha, \hat{S}_1, \hat{S}_2)$ with $\hat{S}_1$ and $\hat{S}_2$ being the estimates of S_1 and S_2.

For $\hat{S}_1$ and $\hat{S}_2$, one can use the NPMLEs of S_1 and S_2 based on the observed univariate interval-censored data

$$\{\, U_i^{(1)}, V_i^{(1)}, \Delta_{1i}^{(1)}, \Delta_{2i}^{(1)} \ ; \ i \,=\, 1, ..., n \,\}$$

and

$$\{\, U_i^{(2)}, V_i^{(2)}, \Delta_{1i}^{(2)}, \Delta_{2i}^{(2)} \ ; \ i \,=\, 1, ..., n \,\} \ ,$$

respectively. These estimates, $\hat{S}_1$ and $\hat{S}_2$, can be obtained from the algorithm described in Section 3.2 or 3.4. Given $\hat{S}_1$ and $\hat{S}_2$, one can estimate α by the solution, say $\hat{\alpha}$, to the pseudo score equation $U(\alpha, \hat{S}_1, \hat{S}_2, \hat{G}_n) \,=\, 0$, where

$$U(\alpha, \hat{S}_1, \hat{S}_2, \hat{G}_n) \,=\, \frac{\partial l(\alpha, \hat{S}_1, \hat{S}_2)}{\partial \alpha} \,=\, n \int \frac{\partial}{\partial \alpha}\, l(\alpha, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \boldsymbol{\delta})\, dG_n(\boldsymbol{x}, \boldsymbol{\delta})$$

and $G_n(\boldsymbol{x}, \boldsymbol{\delta})$ denotes the empirical estimator of $G_\alpha(\boldsymbol{x}, \boldsymbol{\delta})$. It is easily shown that $\hat{\alpha}$ is consistent and the root of equation above can be obtained by standard methods, such as the Newton-Raphson algorithm.

This two-stage procedure was originally proposed by Wang and Ding (2000) and Sun, Wang, and Sun (2006) for cases I and II bivariate interval-censored failure time data, respectively. They show that under some regularity conditions, $n^{1/2}\,(\hat{\alpha} \,-\, \alpha_0)$ converges in distribution to a zero-mean normal random variable as $n \to \infty$, where α_0 denotes the true value of α. For case I bivariate interval-censored data, a simple estimate of the asymptotic variance is given below. For case II bivariate interval-censored data, however, the estimate of the asymptotic variance is quite complicated (Sun, Wang, and Sun, 2006). For this reason, we only consider a bootstrap estimate.

We first consider estimating the variance of $\hat{\alpha}$ for case II bivariate interval-censored data. To obtain a bootstrap estimate, one can draw M independent, simple bootstrap samples of size n with replacement from the observed data $\{\, U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i \ ; \ i \,=\, 1, ..., n \,\}$, where M is a prespecified integer. Then applying the two-stage estimation procedure described above to the bootstrap samples yields M estimates $\{\, \tilde{\alpha}_k \, ; \, k \,=\, 1, ..., M \,\}$ of α. A natural estimate of the variance of $\hat{\alpha}$ is $\hat{\sigma}_\alpha^2 = (M-1)^{-1} \sum_{k=1}^{M} (\tilde{\alpha}_k - \tilde{\alpha})^2$, the sample variance of the $\tilde{\alpha}_k$'s, where $\tilde{\alpha} \,=\, M^{-1} \sum_{k=1}^{M} \tilde{\alpha}_k$. Simulation studies suggest this bootstrap procedure works reasonably well for practical situations with M at least 200 (Sun, Wang, and Sun, 2006).

For case I bivariate interval-censored data, there exists only one monitoring time $U^{(j)}$ for T_j, $j = 1, 2$. Suppose that $U^{(1)} = U^{(2)} = C$ and C is a continuous random variable with density function h. That is, the observation time for T_1 and T_2 is the same, which would be the case, for example, if T_1 and T_2 represent two different failure times from the same subject. In this case, using the variables defined earlier, one has $V^{(1)} = V^{(2)} = C$ and $\Delta_2^{(1)} = \Delta_2^{(2)} = 0$ as mentioned before. Wang and Ding (2000) show that the asymptotic variance of $n^{1/2}\,(\hat{\alpha} - \alpha_0)$ can be consistently estimated by

$$\frac{1}{n-1} \sum_{i=1}^{n} \left[Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \boldsymbol{x}_i, \boldsymbol{\delta}_i) - \bar{Q} \right]^2 ,$$

the sample variance of $\{Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \boldsymbol{x}_i, \boldsymbol{\delta}_i)\ ;\ i = 1, ..., n\}$. In the expression above,

$$Q(\alpha, S_1, S_2, \boldsymbol{x}, \boldsymbol{\delta}) = \frac{\partial}{\partial \alpha} l(\alpha, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \boldsymbol{\delta}) + \frac{g_\alpha(\boldsymbol{x}, \boldsymbol{\delta})}{h(\boldsymbol{x})}$$

$$\times \frac{\partial}{\partial \alpha} \left[\sum_{j=1}^{2} (\, \delta_1^{(j)} - 1 + S_j(x_j)\,) \sum_{\delta_1^{(1)}=0}^{1} \sum_{\delta_1^{(2)}=0}^{1} l_j(\alpha, S_1, S_2, \boldsymbol{x}, \boldsymbol{\delta}) \right]$$

and $\bar{Q} = n^{-1} \sum_{i=1}^{n} Q(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \boldsymbol{x}_i, \boldsymbol{\delta}_i)$, where

$$l_1(\alpha, t_1, t_2, \boldsymbol{x}, \boldsymbol{\delta}) = \frac{\partial l(\alpha, t_1, t_2, \boldsymbol{x}, \boldsymbol{\delta})}{\partial t_1}$$

and

$$l_2(\alpha, t_1, t_2, \boldsymbol{x}, \boldsymbol{\delta}) = \frac{\partial l(\alpha, t_1, t_2, \boldsymbol{x}, \boldsymbol{\delta})}{\partial t_2} .$$

Given $\hat{\alpha}$, one can estimate Kendall's τ from equation (7.2) and apply the delta method to estimate its variance. For example, under the Clayton model, one obtains

$$\hat{\tau} = \frac{\hat{\alpha} - 1}{\hat{\alpha} + 1}$$

with its variance estimated by

$$\hat{\sigma}_\tau^2 = \frac{4\,\hat{\sigma}_\alpha^2}{(\hat{\alpha} + 1)^2} ,$$

where $\hat{\sigma}_\alpha^2$ denotes the estimated variance of $\hat{\alpha}$.

An important application of the estimation procedure discussed above is to test for independence of T_1 and T_2, which implies $\tau = 0$ and in the Clayton model, $\alpha \to 1$. Although one can use either $\hat{\alpha}$ or $\hat{\tau}$ for this and both parameters have similar interpretation, Kendall's τ in general may be more stable than the association parameter α. Of course, one also can use a procedure that is specifically developed to test for independence without assuming the copula model. There exist several such procedures for bivariate right-censored failure time data (Hsu and Prentice, 1996; Oakes, 1982; Shih and Louis, 1995). Ding and Wang (2004) give an approach for bivariate current status data. No procedure seems to be available for case II bivariate interval-censored data.

7.2.3 An Example

We consider bivariate interval-censored failure time data concerning CMV shedding in blood and urine from an AIDS clinical trial on HIV-infected individuals. The data are described in Section 1.2.4 and given in data set I of Appendix A. For the patients in this study, blood and urine samples were supposed to be collected every 12 and 4 weeks, respectively, to test for the presence of CMV. However, as seen in the data set, real sample collection times differ from patient to patient, resulting in interval-censored data for CMV shedding times in both blood and urine. Specifically, for time to CMV shedding in blood, 7 patients have left-censored observations, 174 patients have right-censored observations, and 23 patients have interval-censored observations. For time to CMV shedding in urine, the corresponding numbers of patients are 49, 88 and 67. It is of scientific interest to determine if there is an association within individuals between the time to CMV shedding in blood and the time to CMV shedding in urine.

To assess this relationship, define T_1 and T_2 to be the times to the occurrences of CMV virus in blood and urine, respectively, and assume that they follow the Clayton model. The two-stage estimation procedure gives $\hat{\alpha} = 2.8060$ and with $M = 500$, the estimated standard error is 0.5413. This results in $\hat{\tau} = 0.4746$, and by the delta method, the estimated standard error is 0.0747. As mentioned above, the independence of the CMV shedding times in blood and urine implies $\alpha \to 1$ or $\tau \to 0$. Based on the standard normal distribution, testing $\alpha = 1$ against $\alpha > 1$ gives a p-value of 0.0004. Also based on the standard normal distribution, testing $\tau = 0$ against $\tau > 0$ gives a p-value of less than 0.0001. These results suggest that the CMV shedding times in blood and in urine are significantly correlated for the HIV-infected subjects in the study.

To investigate the effect of the choice of M, we calculated bootstrap estimates $\hat{\sigma}_\alpha$ for different M values and obtained similar results. For example, with $M = 1000$, the estimated standard error is 0.5355.

7.3 Nonparametric Estimation of a Bivariate Distribution Function

Consider a survival study that involves n independent subjects from a homogeneous population with each subject giving rise to two failure times denoted by T_{1i} and T_{2i}, $i = 1, ..., n$. Let $F(t_1, t_2) = P(T_{1i} \leq t_1, T_{2i} \leq t_2)$ denote their joint cumulative distribution function and suppose that only interval-censored failure time data are available. In particular, the observations are

$$\{\, U_i \;=\; (L_{1i}, R_{1i}] \,\times\, (L_{2i}, R_{2i}] \;,\; i = 1, \ldots, n \,\} \;,$$

where $(L_{1i}, R_{1i}]$ and $(L_{2i}, R_{2i}]$ represent the intervals to which T_{1i} and T_{2i} belong, respectively. In other words, the observation on each subject could be a point, line segment (which may be a half-line), or rectangle (which may be a quadrant). These possibilities correspond to the situations where both failure times are observed exactly, one failure time is observed exactly and the other is interval- or right-censored, or both failure times are interval- or right-censored, respectively. If one treats points as rectangles that are degenerate in both dimensions and line segments as rectangles that are degenerate in one dimension, then the observed data consist entirely of rectangles. That is, the observed data are a collection of n rectangles. As before, we use the convention that $(a, a]$ means the single point $\{a\}$.

7.3.1 The Nonparametric Maximum Likelihood Estimator

In this subsection, we study the NPMLE of the distribution function $F(t_1, t_2)$. First, note that the likelihood function of the observed data is proportional to $L(F) \;=\; \prod_{i=1}^n F(U_i)$, where

$$F(U_i) \;=\; F(R_{1i}, R_{2i}) \;-\; F(R_{1i}, L_{2i}) \;-\; F(L_{1i}, R_{2i}) \;+\; F(L_{1i}, L_{2i}) \;.$$

As in the case of univariate interval-censored data discussed in Sections 3.3 and 3.4, it is easy to see that the NPMLE of F has to be discrete. Furthermore, it puts all of the probability mass on the observed rectangles or intersections of observed rectangles. Note that a nonempty intersection of rectangles can be treated as a rectangle too. These rectangles constitute the regions of possible support of the NPMLE of F. Also it is easy to show that the NPMLE can be determined uniquely only up to the probability mass on these regions, but its probability mass can be distributed arbitrarily inside these rectangles. In other words, the likelihood function $L(F)$ is independent of the behavior of F within each of these regions (Betensky and Finkelstein, 1999b; Gentleman and Vandal, 2002).

Let

$$H \;=\; \{\, H_j \;=\; (r_{1j}, s_{1j}] \,\times\, (r_{2j}, s_{2j}] \;,\; j = 1, ..., m \,\}$$

denote the disjoint rectangles that constitute the regions of possible support of the NPMLE of F. The determination of it will be discussed below. Define

$$\alpha_{ij} \;=\; I(H_j \subseteq (L_{1i}, R_{1i}] \,\times\, (L_{2i}, R_{2i}])$$

and

$$p_j \;=\; F(H_j) \;=\; F(s_{1j}, s_{2j}) \;-\; F(r_{1j}, s_{2j}) \;-\; F(s_{1j}, r_{2j}) \;+\; F(r_{1j}, r_{2,j}) \;,$$

$i = 1, ..., n$, $j = 1, ..., m$. Then the likelihood function $L(F)$ can be rewritten as

$$L(\boldsymbol{p}) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} \, p_j \tag{7.3}$$

with $\boldsymbol{p} = (p_1, ..., p_m)'$, and the NPMLE of F is determined by maximizing (7.3) over the p_j's subject to $p_j \geq 0$ and $\sum_{j=1}^{m} p_j = 1$.

Thus, as in the case of univariate interval-censored data, calculation of the maximum likelihood estimator of F involves two steps. First, determine the set of possible support regions, H, and then maximize (7.3) under the constraints $\sum_{j=1}^{m} p_j = 1$ and $p_j \geq 0$ for all j. The second step is the same as the second step for computing the maximum likelihood estimator with univariate interval-censored data. Thus the algorithm described in Section 3.2 or 3.4 can be used. However, as seen below, the first step for the current situation is much more complicated than that for univariate interval-censored data.

For the uniqueness of the p_j's that maximize the likelihood function $L(\boldsymbol{p})$, the conditions given in Section 3.3 for the uniqueness with univariate interval-censored data apply here. Define $A = (\alpha_{ij})$, a $n \times m$ matrix. If the rank of A is m, then the $\boldsymbol{p}$ vector is unique and in this case, the log-likelihood function $\log L(\boldsymbol{p})$ is strictly concave. In general, the log-likelihood function is concave for all cases. The NPMLE may be unique even if the rank of A is smaller than m. In such cases, one could use the sufficient Lagrange-type conditions to check for uniqueness of the NPMLE.

7.3.2 Algorithms for Possible Support Regions

In the following, we discuss three algorithms for finding H. The first algorithm, given by Betensky and Finkelstein (1999b), takes a direct approach and is much simpler in concept, but could be substantially slower than the other two. The second and third algorithms, proposed by Gentleman and Vandal (2001, 2002) and Bogaerts and Lesaffre (2004), respectively, rely on marginal approaches. In particular, the second algorithm makes use of graph theory. In the following, these algorithms are referred to as BF, GV, and BL algorithms, respectively.

7.3.2.1 BF Algorithm

To determine H, we first note that H consists of rectangles that are intersections of observed rectangles such that there are no other rectangles contained within them and each observed rectangle can be expressed as a union of rectangles in H. This is the basic idea behind the BF algorithm, which conducts an iterative, direct search process for all rectangles in H. Before describing the BF algorithm, we need to define an intersection search process. Suppose that there is a set of rectangles that yield K arbitrarily ordered pairwise intersections. Define $\mathcal{M}$ to be a new set of rectangles given as follows. For the

kth ($1 \le k \le K$) intersection, $R_i \cap R_j$, that is a nonempty rectangle, if it is not included in $\mathcal{M}$ yet and also does not properly contain any rectangle that is already in $\mathcal{M}$, add it to $\mathcal{M}$ and discard the two rectangles that give rise to the intersection, i.e., R_i and R_j. If the kth intersection, $R_i \cap R_j$, is empty, check each of the two rectangles, R_i and R_j. If there is a rectangle in $\mathcal{M}$ that is a proper subset of R_i, then discard R_i. Otherwise, add R_i to $\mathcal{M}$. Repeat the same process for R_j. In the following, this search process is referred to as the kept procedure.

Now we are ready to give the BF algorithm for determining H.

Step 0. Apply the kept procedure to the collection of observed rectangles and let $H^{(0)}$ be $\mathcal{M}$, the set of rectangles resulting from their intersections.

Step 1. At the lth iteration, apply the kept procedure to $H^{(l-1)}$ and let $H^{(l)}$ be $\mathcal{M}$, the set of the rectangles resulting from the intersections of all elements in $H^{(l-1)}$.

Step 2. If $H^{(l)} = H^{(l-1)}$, stop and take $H = H^{(l)}$. Otherwise, go back to step 1.

To illustrate this algorithm, consider a simple example from Betensky and Finkelstein (1999b) and Bogaerts and Lesaffre (2004) with $n = 5$ observations represented by the rectangles labeled 1, 2, 3, 4, and 5 in the left panel of Figure 7.1. In the example, observation 4 does not overlap with any other observation. For step 0 of the BF algorithm, we apply the kept procedure to 10 (2 out of 5) intersections and among them, only 4 are nonempty. According to the kept procedure, the resulting $H^{(0)}$ includes the following five rectangles: the original observation 4 and the intersections between observations 1 and 2, 2 and 3, 2 and 5, and 3 and 5. We label the rectangles in $H^{(0)}$ by A, B, C, D, and E, respectively. For the first iteration in step 1, again we need to consider 10 intersections. For the current step, only 3 of 10 are nonempty. The three nonempty intersections are those between C and D, C and E, and

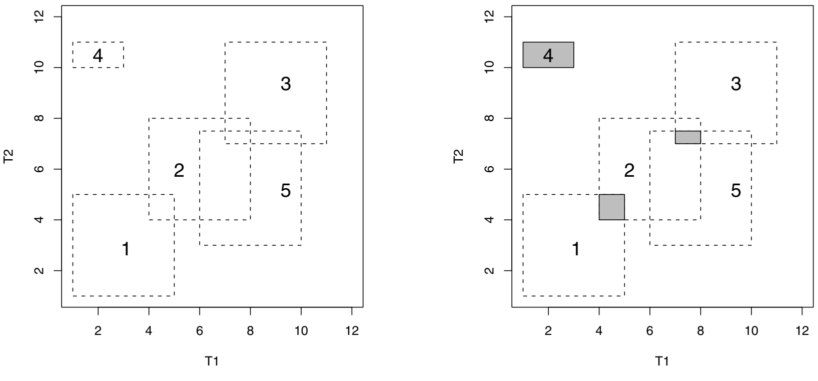

Fig. 7.1. An artificial example for the illustration of the algorithms.

D and E. The last three intersections are identical. Thus $H^{(1)}$ includes only 3 nonoverlapping rectangles, A (observation 4), B (the intersection between observations 1 and 2), and the intersection between C and D (or the intersection given by observations 2, 3 and 5), shaded in the right panel of Figure 7.1. It is apparent that $H = H^{(1)}$.

7.3.2.2 GV Algorithm

For the description of the GV algorithm, we need to define two concepts from graph theory. A *clique* C is a subcollection of the observed rectangles $\{U_i \, , \, i = 1, ..., n\}$ such that each rectangle in C overlaps with every other rectangle in C. A *maximal clique* is a clique that is not a proper subset of any other clique. For the example given in the left panel of Figure 7.1, there exist 3 maximal cliques and they are $\{1, 2\}$, $\{2, 3, 5\}$, and $\{4\}$. By the GV algorithm, finding H is equivalent to finding all maximal cliques for the given observed rectangles. The elements of H are the intersections of all rectangles in each of the maximal cliques. We note that each of the observed rectangles could belong to more than one maximal clique.

To determine all maximal cliques for a set of bivariate interval-censored data, one can use the following three-step procedure.

Step 1. Determine separately the regions of possible support for the marginal distribution functions of T_1 and T_2 based on the univariate interval-censored data $\{ \, (L_{1i}, R_{1i}] \, , \; i = 1, \ldots, n \, \}$ and $\{ \, (L_{2i}, R_{2i}] \, , \; i = 1, \ldots, n \, \}$, respectively. Let them be denoted by $\mathcal{M}_{T_1}$ and $\mathcal{M}_{T_2}$ and recall that they can be obtained by Turnbull's approach given in Section 3.3.

Step 2. For each interval in $\mathcal{M}_{T_1}$, find the subcollection of all the observed rectangles whose projections on the T_1-axis contain it and define $\mathcal{M}_1$ to be the collection of all these subcollections. Perform the same process on each interval in $\mathcal{M}_{T_2}$ and define $\mathcal{M}_2$ to be the collection of the resulting subcollections.

Step 3. To determine, $\mathcal{M}$, the set of all maximal cliques, for each element in $\mathcal{M}_1$, find its intersection with each element in $\mathcal{M}_2$. Recall that elements in $\mathcal{M}_1$ and $\mathcal{M}_2$ are subcollections of observed rectangles. If the intersection, which also is a subcollection of all the observed rectangles, is empty, move to the next intersection. If it is nonempty, check if it already is included in $\mathcal{M}$ or is included as a proper subset of another intersection. If so, move to the next intersection and otherwise, include it in $\mathcal{M}$.

To illustrate the GV algorithm, we again consider the example given in the left panel of Figure 7.1. From step 1, one has $\mathcal{M}_{T_1} = \{(1, 3], (4, 5], (7, 8]\}$ and $\mathcal{M}_{T_2} = \{(4, 5], (7, 7.5], (10, 11]\}$. Step 2 gives $\mathcal{M}_1 = \{\{1, 4\}, \{1, 2\}, \{2, 3, 5\}\}$ and $\mathcal{M}_2 = \{\{1, 2, 5\}, \; \{2, 3, 5\}, \{3, 4\}\}$. For step 3, the nonempty, distinct intersections are $\{1\}$, $\{4\}$, $\{1, 2\}$, $\{2\}$, $\{2, 5\}$, $\{2, 3, 5\}$ and $\{3\}$, resulting in $\mathcal{M} = \{\{4\}, \{1, 2\}, \{2, 3, 5\}\}$. They give the same H as that shown in the right panel of Figure 7.1.

7.3.2.3 BL Algorithm

Step 1 of the GV algorithm starts with one-dimensional situations. The intervals generated by projecting the rectangles in H onto either the T_1-axis or T_2-axis should be part of the intervals of possible support for the marginal distribution of T_1 or T_2, respectively. This observation also motivates the following two-step BL algorithm.

Step 1. First collect all distinct end points from the observed intervals on T_1 only. For each such point, say t_1, find the regions of possible support for the marginal distribution of T_2 based on the data given by intersecting the observed rectangles with the straight line $T_1 = t_1$. Denote by $\mathcal{N}_{1,t_1}$ the collection of these possible support intervals and by $\mathcal{N}_1$ the set of all resulting $\mathcal{N}_{1,t_1}$. Repeat this process on T_2 only to obtain the $\mathcal{N}_{2,t_2}$ and $\mathcal{N}_2$.

Step 2. To find H, for each $\mathcal{N}_{1,t_1}$ in $\mathcal{N}_1$ and each interval, say $(l_2, r_2]$, in $\mathcal{N}_{1,t_1}$, check if there exists an interval with t_1 as either the left or right end point that belongs to both of $\mathcal{N}_{2,l_2}$ and $\mathcal{N}_{2,r_2}$. If not, move to the next interval or next element in $\mathcal{N}_1$. Suppose that such an interval exists, say $(l_1, r_1]$ with $l_1 = t_1$. Then check if $(l_2, r_2] \in \mathcal{N}_{1,r_1}$. If not, move to the next interval or next element in $\mathcal{N}_1$. Otherwise, check if the rectangle $(l_1, r_1] \times (l_2, r_2]$ already belongs to H or includes some rectangle in H as a proper subset. If not, include it in H and otherwise, move to the next interval or next element in $\mathcal{N}_1$ unless all intervals and all elements in $\mathcal{N}_1$ have been checked.

In step 2, one could start with an element in $\mathcal{N}_2$ and arrive at the same H. We apply the BL algorithm to the data given in the left panel of Figure 7.1. For step 1, we note that there exist 9 distinct end points for both T_1 and T_2, respectively. Thus both $\mathcal{N}_1$ and $\mathcal{N}_2$ contain 9 elements, that is, 9 sets of intervals. For example, two end points for T_1 are 3 and 4 and the corresponding sets of possible support intervals are $\mathcal{N}_{1,3} = \{(1,5], (10,11]\}$ and $\mathcal{N}_{1,4} = \{(4,5]\}$, respectively. For step 2, first take the interval $(1,5]$ from $\mathcal{N}_{1,3}$ and check if there exists an interval in both $\mathcal{N}_{2,1}$ and $\mathcal{N}_{2,5}$ that starts or ends with $t_2 = 3$, which is obviously not true. For the next interval $(10,11]$ from $\mathcal{N}_{1,3}$, there does exist an interval that satisfies the needed condition and it is $(1,3]$ belonging to both $\mathcal{N}_{2,10}$ and $\mathcal{N}_{2,11}$. Furthermore, one has that $(10,11] \in \mathcal{N}_{1,1}$. Thus one has one possible support rectangle given by $(1,3] \times (10,11]$, which is the original observation 4. The other two support rectangles are obtained similarly.

7.3.3 An Example

Table 7.1, reproduced from Betensky and Finkelstein (1999b), presents a set of bivariate interval-censored failure time data arising from ACTG 181, the same study discussed in Sections 1.2.4 and 7.2.3. The two failure times of interest here are the time (T_1) to shedding of CMV in the urine and blood and the time (T_2) to colonization of mycobacterium avium complex (MAC) in the sputum and stool. Although the patients in the study were assigned

Table 7.1. Observed rectangles $(L_C, R_C] \times (L_M, R_M]$ in months for CMV shedding and MAC colonization from ACTG 181

L_C	R_C	L_M	R_M	Multiplicities	L_C	R_C	L_M	R_M	Multiplicities
0	3	0	-	3	0	3	2	-	1
0	3	5	-	3	0	6	5	-	1
0	3	8	-	1	0	3	11	-	5
0	3	14	-	5	0	6	14	-	1
2	3	2	-	1	2	3	5	-	1
2	3	8	-	3	2	6	8	-	2
2	6	11	-	3	2	3	14	-	2
2	6	14	-	2	2	6	17	-	1
2	3	20	-	1	5	6	0	-	2
5	9	0	-	1	5	9	8	-	1
5	6	11	-	1	5	9	11	-	2
5	6	14	-	1	5	9	14	-	1
5	6	17	-	1	5	9	17	-	2
8	9	0	-	1	8	12	0	-	2
8	9	8	-	2	8	12	8	-	1
8	12	11	-	3	8	9	14	-	1
8	12	23	-	1	8	9	26	-	1
11	12	0	-	1	11	15	0	-	1
11	15	5	-	1	11	15	14	-	1
11	15	20	-	1	0	-	0	-	6
2	-	0	-	2	5	-	0	-	1
5	-	2	-	2	5	-	5	-	3
5	-	8	-	1	8	-	0	-	2
8	-	8	-	3	8	-	11	-	1
11	-	0	-	5	11	-	5	-	1
11	-	8	-	4	11	-	11	-	10
14	-	0	-	3	14	-	2	-	1
14	-	5	-	1	14	-	8	-	2
14	-	11	-	8	14	-	14	-	9
17	-	0	-	1	17	-	5	-	1
17	-	8	-	1	17	-	11	-	1
17	-	14	-	3	17	-	17	-	6
20	-	14	-	1	-	0	0	-	9
-	0	2	-	3	-	0	5	-	10
-	0	8	-	6	-	0	11	-	8
-	0	14	-	5	-	0	17	-	4
-	0	20	-	1	-	0	0	3	1
5	-	0	6	1	5	-	5	6	1
11	-	0	3	1	11	-	0	6	1
14	-	0	3	1	20	-	14	15	1
2	-	-	0	1	8	-	-	0	1
11	-	-	0	1	0	3	0	6	1
2	6	5	12	1	8	9	8	9	1
-	0	-	0	1					

prescheduled clinic visit times as discussed before, many missed some of the visits and returned with a change in laboratory results for CMV shedding or MAC colonization, thus yielding interval-censored data for T_1 and T_2. The goal here is to estimate the joint distribution function of T_1 and T_2.

The data set consists of 204 patients who were tested for CMV shedding and MAC colonization at least once during the trial, and the time unit is one month. In the table, $(L_C, R_C]$ and $(L_M, R_M]$ represent the observed intervals for T_1 and T_2, respectively. For the data, there exist 68 and 10 interval-censored observations on the times to CMV shedding and MAC colonization, respectively. The numbers of right-censored CMV shedding and MAC colonization times are 89 and 190, respectively, and the remaining observations are left-censored.

To derive the NPMLE of the joint distribution function of T_1 and T_2, we first determine the regions or rectangles of possible support for the joint distribution function. All three algorithms described above suggest that the NPMLE would put probability mass at most on 32 rectangles. That is, for the observed data, there exist 32 maximal cliques and H includes 32 nonoverlapping rectangles. The NPMLE, which is given in Table 7.2, puts positive probability mass on only 13 of the 32 rectangles. Concerning uniqueness of the NPMLE, the rank of A in this example is 31, not 32, suggesting that the NPMLE may not be unique. However, the submatrix of A given by its columns corresponding to 13 zero Lagrange multipliers has rank 13. Thus applying the sufficient Lagrange conditions given in Section 3.3, the NPMLE given in Table 7.2 is unique.

Bogaerts and Lesaffre (2004) and Gentleman and Vandal (2002) also analyzed this data set. In particular, Bogaerts and Lesaffre (2004) compared the

Table 7.2. NPMLE of the joint probability function of times to CMV shedding and MAC colonization

r_{1j}	s_{1j}	r_{2j}	s_{2j}	$\hat{P}((T_1, T_2) \in H_j)$
0	0	0	0	0.014
0	0	20	∞	0.308
2	3	20	∞	0.087
5	6	5	6	0.015
5	6	17	∞	0.063
8	9	8	9	0.010
8	9	27	∞	0.071
11	12	0	0	0.005
11	12	23	∞	0.053
14	15	0	0	0.042
14	15	20	∞	0.022
20	∞	14	15	0.044
20	∞	17	∞	0.266

three algorithms described in Section 7.3.2 and found that for the data, they took 32, 0.07, and 0.17 seconds, respectively, for the determination of H.

7.3.4 Discussion

Concerning the three algorithms for determining H, the BF algorithm takes a direct, two-dimensional approach, whereas the other two apply marginal, univariate approaches. Although both the GV and BL algorithms rely on marginal approaches, they use different search processes. The former begins with univariate intervals and focuses on finding maximal cliques, whereas the latter begins with univariate end points and searches directly for the H_j's.

For many situations, an algorithm will be selected based on speed. To give a simple comparison of the three algorithms, consider two extreme examples. First suppose that there is no overlap between observed rectangles. In this case, the BF algorithm determines that H is simply all the observed rectangles and little effort is needed. In contrast, the fact that H is all the observed rectangles may not be determined right away if one applies either the GV or BL algorithms. For a second example, suppose that all the observed rectangles are different, but every pair of them have a nonempty intersection. Then in the second step of the BF algorithm, one must consider 2 out of n new rectangles, which could lead one to deal with a large number of rectangles in the following steps. However, in this case, it is obvious that there exists only one maximal clique.

In general, for situations where n is small or the overlap between the observed rectangles is light, the BF algorithm seems to be efficient. For large n and substantial overlap of the observed rectangles, the BF algorithm could be too slow (Bogaerts and Lesaffre, 2004) and thus the two algorithms based on marginal approaches should be used. Although the BL algorithm is faster than the GV algorithm in general, the former requires more memory (Bogaerts and Lesaffre, 2004). For generalizations to higher dimensions, the GV algorithm may be more straightforward because of its use of graph theory.

As mentioned in Section 7.3.1, given the rectangles of possible support, H, the determination of the NPMLE, the maximization of $L(F)$, is similar to that for univariate interval-censored data. However, the estimation problem discussed in this chapter is much more complicated, and the estimators can be more problematic than that in the case of univariate interval-censored data. For example, the study of theoretical properties such as consistency is more challenging, and the NPMLE may more often exhibit undesirable properties such as lack of consistency and non-uniqueness. This also is true even for bivariate right-censored failure time data.

Because of possible non-uniqueness, inconsistency, and other problems of the NPMLE, research has been conducted to repair or modify the NPMLE or to impose some assumptions to overcome these problems (Kalbfleisch and Prentice, 2002). For example, van der Laan (1996) gave a repaired NPMLE based on a reduced data set that could perform better than the original

NPMLE for bivariate right-censored failure time data. Kang and Koehler (1998) and Yu, Wong, and He (2000) discussed these problems and proposed some estimators for bivariate interval-censored data. Maathuis (2005) also gave a reduction algorithm. To use a smoothing estimator provides another approach (He and Lawless, 2003).

An alternative to nonparametric maximum likelihood estimation of F is to use the so-called plus-in approach, often used for bivariate right-censored failure time data (Kalbfleisch and Prentice, 2002; Prentice and Cai, 1992, Lin and Ying, 1993). In this approach, a bivariate survival function is expressed as a function of some other quantities that can be more easily estimated. For example, Prentice and Cai (1992) express the joint survival function of T_1 and T_2 as

$$S(t_1,t_2) \, = \, S_1(t_1)\, S_2(t_2) \left[1 + \int_0^{t_1} \int_0^{t_2} [S_1(s_1)S_2(s_2)]^{-1}\, C(ds_1, ds_2) \right] \, ,$$

where S_1 and S_2 denote the marginal survival functions as before and $C(t_1, t_2)$ is the covariance function of T_1 and T_2. This suggests that one can estimate $S(t_1, t_2)$ by using estimates of S_1, S_2 and $C(t_1, t_2)$ in the expression for $S(t_1, t_2)$. Alternatively, if model (7.1) holds, then estimation of $S(t_1, t_2)$ reduces to estimation of the marginal survival functions $S_1(t_1)$ and $S_2(t_2)$ and the association parameter α.

7.4 Regression Analysis with the Grouped Proportional Hazards Model

This section considers regression analysis of discrete bivariate interval-censored failure time data. As before, let T_1 and T_2 denote two possibly correlated failure time variables. Suppose that the time axis is divided into m time intervals or the possible probability mass points for T_1 and T_2 are $s_1 < ... < s_m$. If the probability mass points for T_1 and T_2 are different, then an inference procedure like the one below can be developed or one can use the procedure given below using the union of the probability mass points for T_1 and T_2. It is straightforward to generalize the methods discussed here to interval-censored data with dimension greater than two.

First we describe three marginal grouped PH models that are commonly used for the analysis of multivariate failure time data. Some inference procedures are then presented using marginal approaches (Goggins and Finkelstein, 2000; Kim and Xue, 2002), which have the advantage of leaving the association between failure time variables arbitrary. An illustrative example follows.

7.4.1 The Grouped Proportional Hazards Model

Let $S_j(t; \boldsymbol{Z}) \, = \, P(T_j > t | \boldsymbol{Z})$ denote the marginal survival function of T_j given a vector of covariates $\boldsymbol{Z}$, $j \, = \, 1, 2$. For regression analysis, one can apply one of the following three marginal regression models:

$$S_{j,k}(\boldsymbol{Z}) = S_j(s_k; \boldsymbol{Z}) = (q_1 \cdots q_k)^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})} \,, \tag{7.4}$$

$$S_{j,k}(\boldsymbol{Z}) = S_j(s_k; \boldsymbol{Z}) = (q_{j,1} \cdots q_{jk})^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})} \,, \tag{7.5}$$

and

$$S_{j,k}(\boldsymbol{Z}) = S_j(s_k; \boldsymbol{Z}) = (q_{j,1} \cdots q_{j,k})^{\exp(\boldsymbol{Z}'\boldsymbol{\beta}_j)} \,. \tag{7.6}$$

In (7.4), (7.5) and (7.6),

$$q_{j,k} = P(T_j > s_k \,|\, T_j > s_{k-1}, \boldsymbol{Z} = 0)$$

with $s_0 = 0$, q_k denotes $q_{j,k}$ if T_1 and T_2 have the same marginal distribution, and $\boldsymbol{\beta}$ or the $\boldsymbol{\beta}_j$'s denote the vector or vectors of regression parameters, $k = 1, ..., m$, $j = 1, 2$.

Among these three models, model (7.4) is the simplest and assumes that the baseline survival functions for T_1 and T_2 are the same as well as the covariates' effects on them. In other words, it imposes the same marginal distribution on T_1 and T_2. This could be appropriate for situations such as eye studies. In contrast, both models (7.5) and (7.6) allow the baseline survival functions for T_1 and T_2 to be different. Model (7.5) applies to situations where the effects of covariates on T_1 and T_2 can be reasonably assumed to be identical, while model (7.6) should be used if the effects of covariates on T_1 and T_2 may be different. Goggins and Finkelstein (2000) and Kim and Xue (2002) considered the analysis of general multivariate interval-censored data using models (7.5) and (7.6), respectively.

Under these models, the corresponding density functions are

$$f_{j,k}(\boldsymbol{Z}) = (q_1 \cdots q_{k-1})^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})} \, [1 - q_k^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}] \,,$$

$$f_{j,k}(\boldsymbol{Z}) = (q_{j,1} \cdots q_{j,k-1})^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})} \, [1 - q_{j,k}^{\exp(\boldsymbol{Z}'\boldsymbol{\beta})}]$$

and

$$f_{j,k}(\boldsymbol{Z}) = (q_{j,1} \cdots q_{j,k-1})^{\exp(\boldsymbol{Z}'\boldsymbol{\beta}_j)} \, [1 - q_{j,k}^{\exp(\boldsymbol{Z}'\boldsymbol{\beta}_j)}] \,,$$

respectively. Define $\gamma_k = \log[-\log(q_k)]$ and $\boldsymbol{\theta} = (\gamma_1, ..., \gamma_{m-1}, \boldsymbol{\beta}')$ for model (7.4), $\gamma_{j,k} = \log[-\log(q_{j,k})]$ and $\boldsymbol{\theta} = (\gamma_{1,1}, \gamma_{1,2}, ..., \gamma_{2,m-1}, \boldsymbol{\beta}')$ for model (7.5), or $\boldsymbol{\theta}_j = (\gamma_{j,1}, ..., \gamma_{j,m-1}, \boldsymbol{\beta}_j')$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ for model (7.6). Then the density functions can be rewritten as

$$f_{j,k}(\boldsymbol{Z}, \boldsymbol{\theta}) = e^{-(e^{\gamma_1} + ... + e^{\gamma_{k-1}}) \, \exp(\boldsymbol{Z}'\boldsymbol{\beta})} \, (1 - e^{-\exp(\gamma_k + \boldsymbol{Z}'\boldsymbol{\beta})}) \,,$$

$$f_{j,k}(\boldsymbol{Z}, \boldsymbol{\theta}) = e^{-(e^{\gamma_{j,1}} + ... + e^{\gamma_{j,k-1}}) \, \exp(\boldsymbol{Z}'\boldsymbol{\beta})} \, (1 - e^{-\exp(\gamma_{j,k} + \boldsymbol{Z}'\boldsymbol{\beta})}) \,,$$

and

$$f_{j,k}(\boldsymbol{Z}, \boldsymbol{\theta}) = e^{-(e^{\gamma_{j,1}} + ... + e^{\gamma_{j,k-1}}) \, \exp(\boldsymbol{Z}'\boldsymbol{\beta}_j)} \, (1 - e^{-\exp(\gamma_{j,k} + \boldsymbol{Z}'\boldsymbol{\beta}_j)}) \,,$$

respectively. This reparameterization is commonly used to remove the range restriction on the parameter space and to improve the convergence rate of the iterative estimation procedure used to solve the score functions.

The marginal PH model is often used for the analysis of multivariate right-censored failure time data (Guo and Lin, 1994; Lin, 1994; Wei et al., 1989). To derive an estimation procedure for the regression parameters with such data, T_1 and T_2 are usually assumed to be independent, which is commonly referred to as the working independence assumption. In the following, we adopt the same approach for estimation of the parameters in models (7.4) - (7.6).

7.4.2 Inference Procedures

As in the previous section, suppose that the observed data are

$$\{\, U_i \,=\, (L_{1i}, R_{1i}] \,\times\, (L_{2i}, R_{2i}]\,,\, \boldsymbol{Z}_i\,,\, i = 1, \ldots, n \,\}\,,$$

where $(L_{1i}, R_{1i}]$ and $(L_{2i}, R_{2i}]$ are the intervals, which include the survival times, T_{1i} and T_{2i}, associated with subject i, respectively. Define $\alpha^j_{i,k} = I(s_j \epsilon (L_{ji}, R_{ji}])$, $i = 1, ..., n$, $k = 1, ..., m$, $j = 1, 2$. Then the marginal likelihood for T_j is proportional to

$$L_j(\boldsymbol{\theta}) \,=\, \prod_{i=1}^{n} \sum_{k=1}^{m} \alpha^j_{i,k}\, f_{j,k}(\boldsymbol{Z}_i, \boldsymbol{\theta})\,,$$

$j = 1, 2.$

Under model (7.4) or (7.5), using the working independence assumption, the log likelihood function is

$$l(\boldsymbol{\theta}) \,=\, \sum_{j=1}^{2} \log\left[\, L_j(\boldsymbol{\theta})\,\right]$$

and one estimates $\boldsymbol{\theta}$ by maximizing $l(\boldsymbol{\theta})$. Under model (7.6), one also can maximize $l(\boldsymbol{\theta})$. In this case, the working independence assumption is not required, because one can maximize $L_1(\boldsymbol{\theta})$ and $L_2(\boldsymbol{\theta})$ separately. This estimator of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, can be obtained by, for example, the Newton-Raphson algorithm. The functions $L(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta})\, L_2(\boldsymbol{\theta})$ and $L_j(\boldsymbol{\theta})$ are similar to the likelihood functions considered in Section 6.2. Thus the formulas given there for the first and second derivatives of the log likelihood function can be used here with minor modifications along with the Newton-Raphson algorithm to solve the resulting score functions.

Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$ and define

$$\boldsymbol{U}_{i,j}(\boldsymbol{\theta}) \,=\, \frac{\partial l_{i,j}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \,=\, \frac{\partial}{\partial \boldsymbol{\theta}} \log\left[\sum_{k=1}^{m} \alpha^j_{i,k}\, f_{j,k}(\boldsymbol{Z}_i, \boldsymbol{\theta})\right]$$

and

$$I(\boldsymbol{\theta}) \;=\; -\,\frac{1}{n}\,\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\,,$$

$i = 1, ..., n$, $j = 1, 2$. Under some regularity conditions, it can be shown that $\hat{\boldsymbol{\theta}}$ is consistent as long as the marginal models, (7.4) to (7.6), are correctly specified. Furthermore, $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to the multivariate normal random vector with mean zero and covariance matrix that can be consistently estimated by $I^{-1}(\hat{\boldsymbol{\theta}})\,D(\hat{\boldsymbol{\theta}})\,I^{-1}(\hat{\boldsymbol{\theta}})$, where

$$D(\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j_1=1}^{2}\sum_{j_2=2}^{2}\boldsymbol{U}_{i,j_1}(\boldsymbol{\theta})\,\boldsymbol{U}'_{i,j_2}(\boldsymbol{\theta})$$

(Guo and Lin, 1994).

For estimation of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under model (7.6), as mentioned above, one can simply maximize the log likelihood functions $l_1(\boldsymbol{\theta}_1) = \log L_1(\boldsymbol{\theta}_1)$ and $l_2(\boldsymbol{\theta}_2) = \log L_2(\boldsymbol{\theta}_2)$, respectively. Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ denote these estimators of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Then, instead of using the general variance estimate given above, Kim and Xue (2002) suggest that one can simply estimate the asymptotic covariance matrix of $\sqrt{n}\,(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_{j,0})$ by $I_j^{-1}(\hat{\boldsymbol{\theta}}_j)$, where

$$I_j(\boldsymbol{\theta}_j) \;=\; -\,\frac{1}{n}\,\frac{\partial^2 l_j(\boldsymbol{\theta}_j)}{\partial\boldsymbol{\theta}_j\,\partial\boldsymbol{\theta}'_j}$$

and $\boldsymbol{\theta}_{j,0}$ denotes the true value of $\boldsymbol{\theta}_j$, $j = 1, 2$. In this case, one also can easily estimate the covariance between $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$.

Models (7.4) to (7.6) assume that there exists a single set of covariates whose effects on both T_1 and T_2 are of interest. In practice, the covariates that affect T_1 and T_2 may be different. Suppose that for all subjects, the covariates that affect T_1 and T_2 are given by two $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, which may be completely different or share some common elements. In this case, for models (7.4) and (7.5), we define a new vector of covariates, say $\boldsymbol{Z}^*$, by combining $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, and for each study subject, replace $\boldsymbol{Z}$ by $\boldsymbol{Z}^*$. For model (7.6), we can simply replace $\boldsymbol{Z}$ by $\boldsymbol{Z}_j$. Then the inference procedures given above can be applied.

7.4.3 An Example

To illustrate the inference procedures described in Section 7.4.2, we consider the binary interval-censored failure time data discussed in Sections 1.2.4 and 7.2.3. In that study, it is of interest to estimate the effects of the baseline HIV infection stage on CMV shedding in blood and urine. In other words, it is of interest to assess the association between baseline CD4 cell counts and CMV shedding in blood and urine. A simple way to measure this association is to fit the PH model separately to the data on CMV shedding in blood and urine,

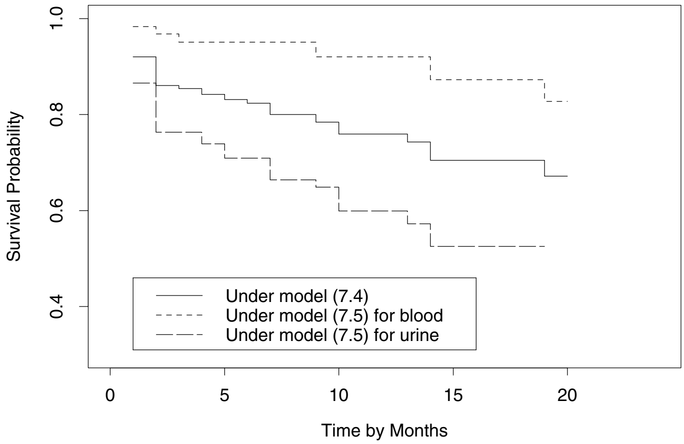

Fig. 7.2. Estimates of baseline survival functions for blood and urine CMV shedding.

respectively, using the methods given in Section 6.2. Because, as shown in Section 7.2.3, the two shedding times are significantly related, it seems more appropriate to fit the PH model to both CMV shedding times together by using the approach of the last subsection.

Define T_1 and T_2 to be the times to CMV shedding in blood and urine, respectively, and $Z = 1$ if the baseline CD4 cell count is less than 75 (cells/μl) and $Z = 0$ otherwise. First we assume that T_1 and T_2 share a common baseline survival function and the covariate effect is the same for T_1 and T_2. That is, model (7.4) is true. The inference procedure described above gives $\hat{\beta} = 0.8446$ with an estimated standard error of 0.1729, yielding a p-value close to zero for testing $\beta = 0$. This indicates that the patients with baseline CD4 cell count below 75 (cells/μl) have significantly higher risk of CMV shedding in blood or urine than those with baseline CD4 cell count above 75 (cells/μl).

Allowing T_1 and T_2 to have different baseline survival functions, but assuming the covariate effects are the same, the inference procedure under model (7.5) gives $\hat{\beta} = 0.9503$ with estimated standard error being 0.1932. Again, this corresponds with a p-value close to zero, and the result is similar to that under model (7.4). Figure 7.2 presents the estimates of the baseline survival functions of T_1 and T_2 obtained under models (7.4) and (7.5), respectively. It suggests that T_1 and T_2 seem to have different baseline survival functions and that the CMV shedding in urine seems to occur much earlier than that in blood. Because T_1 and T_2 appear to have different survival functions, model (7.5) may be more appropriate than model (7.4) for the data set. If one fits model (7.6) to the observed data, the method gives $\hat{\beta}_1 = 1.1997$ and $\hat{\beta}_2 = 0.8893$

with estimated standard errors of 0.4159 and 0.1993, respectively. Note that this is equivalent to fitting the PH model to T_1 and T_2 separately. The results obtained here suggest that the application of separate analyses could result in much less significant covariate effects. In other words, the approach given above for the simultaneous analysis of two correlated, interval-censored failure times could be more powerful than the separate analyses.

7.5 Bibliography, Discussion, and Remarks

The literature on bivariate or multivariate interval-censored failure time data is relatively limited compared with that on other topics discussed in the previous chapters. The existing references mainly focus on the three areas discussed in the previous three sections: inference about the association parameter between two correlated survival variables, nonparametric estimation of joint distribution or survival functions, and regression analysis. As discussed above, for the first area, the available references include Betensky and Finkelstein (1999a), Ding and Wang (2004), Sun, Wang, and Sun (2006), and Wang and Ding (2000), and the authors who considered the second area include Betensky and Finkelstein (1999b), Bogaerts and Lesaffre (2004), Gentleman and Vandal (2001, 2002), Kang and Koehler (1998), and Yu, Wong, and He (2000). For the regression analysis problem, the references include Bogaerts et al. (2002), Goggins and Finkelstein (2000), He and Lawless (2003), and Kim and Xue (2002). In addition, Jewell et al. (2005) discussed estimation of smooth functionals of the marginal distribution functions for current status data.

There are many open problems and issues that need to be studied for the analysis of bivariate interval-censored failure time data. As with bivariate right-censored data, bivariate interval-censored data are much more challenging than their univariate counterparts. In addition to the problems and issues that exist for univariate interval-censored data, an analysis of bivariate data has to address the problems arising because the two failure time variables may be correlated.

In the analysis of bivariate failure time data, nonparametric estimation of a joint survival function and regression analysis are two of the most basic components. Many issues remain untouched even when only right censoring exists (Hougaard, 2000) and of course, interval censoring makes them even more challenging. For example, there exists little research on the properties of the NPMLE discussed in Section 7.3. Of course, one could use other estimators of a joint distribution function rather than the NPMLE. For instance, in both nonparametric estimation and regression analysis in the presence of interval censoring, one could use the conditional or frailty approach (Kalbfleisch and Prentice, 2002; Klein and Moeschberger, 2003). For this approach, suppose that there exists a shared and unobserved frailty W, a positive random variable with mean one, and conditional on $W = w$, T_1 and T_2 are independent with the hazard function of T_j given by

$$w\,\lambda_j(t)\;,\; j\;=\;1,2. \tag{7.7}$$

Then the joint survival function of T_1 and T_2 has the form

$$S(t_1,t_2)\;=\;E\,\exp\left[\,-\,W\,(\,\Lambda_1(t_1)+\Lambda_2(t_2)\,)\,\right]\;=\;L\left[\,\Lambda_1(t_1)+\Lambda_2(t_2)\,\right]\;,$$

where $\Lambda_j(t)\;=\;\int_0^t\lambda_j(s)ds$ denotes the cumulative hazard function of T_j and L is the Laplace transformation of the distribution of W. Under model (7.7), T_1 and T_2 have a nonnegative association and they are independent if the variance of W is zero. That is, W has a degenerate distribution.

Gamma distributions are often used to model W. In this case, $S(t_1,t_2)$ is given by the copula model with C_α belonging to the Clayton family discussed in Section 7.2, and model (7.7) is called the gamma frailty model. Other commonly used distributions for W include the log normal distributions, the positive stable distributions and the inverse Gaussian distributions. For regression analysis, one can replace model (7.7) by

$$w\,\lambda_j(t)\;\exp(\boldsymbol{Z}'\boldsymbol{\beta})\;, \tag{7.8}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{Z}$ are defined as in Section 7.4. Compared with the marginal approach, one advantage of this frailty model approach is that it directly models the correlation of T_1 and T_2. However, one of its disadvantages is that the marginal hazard functions do not follow the PH model. Also the regression parameters do not have the same interpretations as under the marginal PH model.

In this chapter, only the grouped PH model is considered for regression analysis of bivariate interval-censored data. However, the marginal inference approach can be used with other regression models, such as the continuous, i.e., the usual PH model, the additive hazards model, the proportional odds model or others considered in Chapter 5 or 6 and apply the marginal approach for inference. For bivariate right-censored data with the PH model, the marginal approach yields an inference procedure that only involves regression parameters, which is an advantage. This is not the case for bivariate interval-censored data. For this reason, the full likelihood approach may be a better choice because both involve estimation of baseline hazard functions and regression parameters. However, a drawback of the full likelihood approach is that it needs more detailed modeling specifications, such as model (7.7) or (7.8), and could be more complicated, but it could be more efficient.

Given the complexity of fully nonparametric and semiparametric inferences for bivariate interval-censored failure time data, an alternative approach is to impose a smoothness or even a piecewise constant assumption on, for example, the hazard function as discussed in Sections 2.5 and 3.5. The approach is easier to implement than the nonparametric and semiparametric ones, but more flexible than parametric approaches. He and Lawless (2003) investigated this with using piecewise constant and spline specifications for baseline hazard functions in regression analysis of bivariate failure time data.

It is of practical importance to extend the inferences discussed in this chapter to multivariate interval-censored failure time data with dimension greater than two. In this case, there are several possibly correlated failure time variables. For examples of multivariate failure time data, readers are referred to Hougaard (2000). Some of the inference procedures presented in this chapter, such as those in Section 7.4, can be easily generalized to multivariate interval-censored data, whereas the generalization of the others, such as those in Section 7.3, is not straightforward.

This page intentionally blank

8

Analysis of Doubly Censored Data

8.1 Introduction

As discussed in Section 1.3, doubly censored data occur in studies that consist of two related events with one followed by the other. A typical example is given by a disease progression study in which the onset of the disease is caused or preceded by certain virus infection. In these situations, three variables are present, and they are time to infection, time between infection and the onset of the disease, and time to the onset of the disease. It is apparent that one only needs to know two of the three variables. If the variable of interest is the time to infection or the time to the onset of the disease, in general, one only needs to analyze the variable of interest without the need of dealing with the other two variables.

By doubly censored data, we usually mean that the variable of interest is the time between infection and the onset of certain disease such as AIDS latency time in AIDS studies. In this case, one has to deal with two of the three variables together as seen below. In other words, one cannot transform doubly censored data into general interval-censored data unless the time to the first event can be observed exactly. Thus different approaches are required for their analyses. Of course, one can approach the analysis of doubly censored data from the point of bivariate data analysis, but the special structure of doubly censored data makes general bivariate data analysis methods inappropriate.

For the analysis of doubly censored data, it is usually convenient to deal with the first two variables, the time to infection and the time between infection and the onset of the disease. The involvement of two variables also means that one has to face two censoring mechanisms. In addition, sometimes one also may need to consider the relationship of the first two variables, although it seems reasonable for most situations to assume that they are independent. In the following, we focus on this independent situation, and remarks on dependent situations are given in Section 8.6.

In Section 8.2, we consider one-sample problem, and several procedures are described for nonparametric estimation of distribution functions for doubly censored data. In addition, we discuss situations that involve truncation as well as double censoring. Section 8.3 deals with semiparametric regression analysis of doubly censored data. Methods are described for inference about regression parameters in both the PH and additive hazards models. The topic of Section 8.4 is nonparametric comparison of survival functions when only doubly censored data are available. In Section 8.5, two illustrative examples are provided for the discussed approaches. Section 8.6 contains bibliographic notes about the analysis of doubly censored data and some discussion regarding several inference approaches and issues that are not treated in this chapter.

For the discussion below, without loss of generality, we assume that the observation on the first variable, the time to infection, is either exact or truly interval-censored, but not right-censored. The observation on the third variable, the onset of the disease, can be either interval- or right-censored.

8.2 Nonparametric Estimation of Distribution Functions

Consider a study that involves n independent subjects from a homogeneous population and gives doubly censored failure time data. For subject i, let X_i and S_i denote the times of the occurrences of two related events with $X_i \leq S_i$, respectively, $i = 1, ..., n$. Also, let $T_i = S_i - X_i$ denote the survival time of interest. In this section, we assume that all of X_i, S_i, and T_i are discrete random variables.

Let $u_1 < ... < u_r$ denote the possible mass points for the X_i's and $v_1 < ... < v_s$ the possible mass points for the T_i's. Define $w_j = Pr(X_i = u_j)$ and $f_k = Pr(T_i = v_k)$, $j = 1, ..., r$, $k = 1, ..., s$. Then $w = \{ w_j \}$ and $f = \{ f_k \}$ with $\sum_{j=1}^{r} w_j = 1$ and $\sum_{k=1}^{s} f_k = 1$ are the probability functions of the X_i's and T_i's, respectively. The goal is to estimate f as well as w.

In the following, we discuss three algorithms that use the self-consistency idea discussed in Section 3.4.1. First we consider a procedure, originally proposed in De Gruttola and Lagakos (1989), that is based on the maximum likelihood approach and is a generalization of the self-consistency algorithm given in Section 3.4.1. The second method, from Gómez and Lagakos (1994) and a two-step procedure, is a simplification of the first procedure and provides a trade-off between efficiency and complexity with respect to the first algorithm. The last method is a conditional likelihood-based approach due to Sun (1997b). It can also be regarded as a generalization of the self-consistency algorithm given in Section 3.4.1 in that it allows truncation as well as double censoring.

8.2.1 A Maximum Likelihood Approach

Suppose that observed data have the form

$$\{\,(L_i, R_i]\,,\,(U_i, V_i]\,,\, i\,=\,1,...,n\,\}$$

such that $X_i \in (L_i, R_i]$ and $S_i \in (U_i, V_i]$. Define $\alpha^i_{jk} = I(L_i < u_j \leq R_i\,,\, U_i < u_j + v_k \leq V_i)$, $j = 1, ..., r$, $k = 1, ..., s$. Then the full likelihood function has the form

$$L_F(w, f) = \prod_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{s} \alpha^i_{jk}\, w_j\, f_k\,. \tag{8.1}$$

To estimate w and f, one can maximize the full likelihood function $L_F(w, f)$ or solve the score equations given by L_F. For this, define

$$I^i_{jk} = \frac{\alpha^i_{jk}\, w_j\, f_k}{\sum_{l=1}^{r} \sum_{m=1}^{s} \alpha^i_{lm}\, w_l\, f_m}\,, \tag{8.2}$$

$$w^*_j = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{s} I^i_{jk}\,\,,\,\, f^*_k = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} I^i_{jk}\,, \tag{8.3}$$

$j = 1, ..., r$, $k = 1, ..., s$. Note that the quantity I^i_{jk} is the conditional expectation of the event $X_i = u_j$ and $T_i = v_k$ given w and f. Thus $\sum_{i=1}^n I^i_{jk}$ provides a natural estimate of the number of subjects with $X_i = u_j$ and $T_i = v_k$ if w and f are known. This motivates the following self-consistency algorithm given by De Gruttola and Lagakos (1989) for estimation of w and f.

Step 1. Choose starting values for w and f.
Step 2. Computer the I^i_{jk}'s given in equation (8.2) and determine the updated estimates w^*_j's and f^*_k's of the w_j's and f_k's given by equation (8.3).
Step 3. Repeat step 2 until the desired convergence occurs.

Let the $\hat{w}_j$'s and $\hat{f}_k$'s denote the estimators of w and f given by this algorithm. Then the cumulative distribution functions of the X_i's and T_i's can be estimated, respectively, by

$$\hat{H}(x) = \sum_{j:u_j \leq x} \hat{w}_j\,\,,\,\, \hat{F}(t) = \sum_{k:v_k \leq t} \hat{f}_k\,.$$

Note that if the X_i's are observed exactly, the algorithm given above is equivalent to the self-consistency algorithm given in Section 3.4 for interval-censored failure time data. In this case, alternatively, one can apply the algorithm given in Section 3.4 to the data $\{\,(U_i - X_i, V_i - X_i]\,\}$ for estimation of f. Of course, the cumulative distribution function of the X_i's can be separately and easily estimated by its empirical estimate.

To implement the algorithm, one needs to choose initial estimators. For this , one can apply the middle point imputation approach described in Section 2.4 to the observed doubly censored data and use the resulting maximum likelihood estimators as initial estimators. Alternatively, one can estimate w and f separately by using the observed interval-censored data on the X_i's and

T_i's, respectively, or apply the middle point imputation approach to these two sets of interval-censored data. To check the convergence of the algorithm, one can use the criterion

$$\sum_{j=1}^{r} |w_j^{*\,(l)} - w_j^{*\,(l-1)}| + \sum_{k=1}^{s} |f_k^{*\,l} - f_k^{*\,(l-1)}| \le \epsilon$$

for a prefixed $\epsilon > 0$, where the $w_j^{*\,(l)}$'s and $f_k^{*\,(l)}$'s denote the estimators of w and f at the lth iteration, respectively.

Note that the estimators $\hat{w}_j$'s and $\hat{f}_k$'s may not maximize L_F or be unique for certain situations. De Gruttola and Lagakos (1989) show that these estimators are either a saddle point or local maximum of the likelihood function L_F. One way to distinguish the maximum likelihood estimator from a saddle point is to examine the matrix evaluated at the final estimators of negative second derivatives of the logarithm of L_F with respect to w and f. If the eigenvalues of the matrix are positive, then the estimator is a local or globe maximum, and if they are both positive and negative, the estimator is a saddle point. However, for most situations, the estimators obtained using the algorithm given here are unique and converge in probability to the true parameters as n increases.

In practice, the nonuniqueness tends to occur when the time points u_j's and v_k's are defined too finely or too many. Use of the less number of time points and thus the less number of parameters are more likely to yield a unique estimator. It also reduces the convergence time of the algorithm. However, too few number of time points would prevent the estimation of some important change patterns of a distribution function.

8.2.2 A Two-step Approach

This subsection considers another algorithm and it estimates w and f separately. For simplicity, assume that observed data are given as in the previous subsection, but with $U_i = V_i$ or $V_i = \infty$, $i = 1, ..., n$. That is, for the S_i's, one has right-censored data instead of interval-censored data. The method can be easily generalized to interval-censored data situations.

To motivate the algorithm, note that if one is interested only in estimating w, it is natural to employ the following marginal likelihood

$$L_m(w) = \prod_{i=1}^{n} \sum_{j\,:\,L_i < u_j \le R_i} w_j$$

based only on interval-censored data on the X_i's. For estimation of f with known w, one can maximize the full likelihood function L_F given in (8.1), which can be rewritten as

$$L_F(f|w) = \prod_{i=1}^{n} \left\{ \sum_{j,k: L_i < u_j \le R_i, u_j + v_k = V_i} w_j \, f_k \right\}^{\delta_i^1 \, \delta_i^2}$$

$$\times \left\{ \sum_{j \, : \, L_i < u_j \le R_i} \; \sum_{k: u_j + v_k > U_i} w_j \, f_k \right\}^{\delta_i^1 \, (1 - \delta_i^2)} . \qquad (8.4)$$

Here $\delta_i^1 = I(R_i < \infty)$ and $\delta_i^2 = I(U_i = V_i)$, $i = 1, ..., n$. Note that the full likelihood function in (8.4) implies that there do not exist right-censored observations on the X_i's as assumed above. Otherwise, it would involve another factor accounting for right-censored X_i. These arguments motivate the following two-step algorithm due to Gómez and Lagakos (1994).

Step 1. Estimate w using the nonparametric marginal maximum likelihood estimator from L_m, which can be obtained by the methods given in Section 3.4.

Step 2. Given $w_j = \hat{w}_j$ from step 1 and the current estimator of f_k, say $\hat{f}_k^{(l)}$, calculate the updated estimator of f_k given by

$$\hat{f}_k^{(l+1)} = \frac{1}{n - m} \left[N_1^{(l)}(k) + N_2^{(l)}(k) \right] ,$$

where

$$N_1^{(l)}(k) = \sum_{i=1}^{n} \delta_i^1 \, \delta_i^2 \, \frac{\phi_{ij} \, \hat{w}_{j(u_j = V_i - v_k)} \, \hat{f}_k^{(l)}}{\sum_{j=1}^{r} \phi_{ij} \, \hat{w}_j \, \hat{f}_{j(v_j = V_i - u_j)}^{(l)}} ,$$

$$N_2^{(l)}(k) = \sum_{i=1}^{n} \delta_i^1 \, (1 - \delta_i^2) \, \frac{\sum_{j=1}^{r} \phi_{ij} \, I(U_i < u_j + v_k) \, \hat{w}_j \, \hat{f}_k^{(l)}}{\sum_{j=1}^{r} \phi_{ij} \, \hat{w}_j \, \sum_{h: v_h > U_i - u_j} \hat{f}_h^{(l)}} ,$$

$m = \sum_{i=1}^{n} (1 - \delta_i^1)$, where $\phi_{ij} = I(u_j \in (L_i, R_i])$.

Step 3. Repeat step 2 until the desired convergence occurs.

The main idea behind this two-step algorithm is to replace doubly censored data by two separate sets of interval-censored data. Its main advantage over the maximum likelihood approach given in the previous subsection is that it can be more easily implemented and make use of the algorithms discussed in Section 3.4 for general interval-censored data. Also it does not have the saddle point or nonuniqueness problem as the maximum likelihood approach. But it could be less efficient than the latter.

Note that for estimation of f, more conveniently, sometimes one may want to transform the observed doubly censored data into a single set of case II interval-censored data as $\{ \, (U_i - R_i, V_i - L_i] \, \}$. Although this would avoid estimation of w and make it possible to employ the approaches discussed in Section 3.4, it is invalid. To see this, note that the resulting likelihood function is given by

$$\prod_{i=1}^{n} \sum_{k=1}^{s} I(U_i - R_i < v_s \le U_i - L_i) \, f_k ,$$

which differs from the likelihood function L_F given in (8.4) and would yield biased estimators.

8.2.3 A Conditional Likelihood-based Approach

Truncation may exist in a disease progression study, and here we consider situations where for subject i, the truncation is characterized by an interval $(B_i^1, B_i^2]$ for S_i, $i = 1, ..., n$. In other words, subject i is included in the study only if S_i is within the interval $(B_i^1, B_i^2]$. Consequently, $(U_i, V_i] \subseteq (B_i^1, B_i^2]$ and one has truncated and doubly censored data on the T_i's given by

$$\{ (L_i, R_i] \, , \, (U_i, V_i] \, , \, (B_i^1, B_i^2] \, , \, i = 1, ..., n \} \, .$$

If $B_i^1 = 0$ and $B_i^2 = \infty$, one then has doubly censored failure time data. In the following, we will assume that $\sum_{k \in \cup_i B_i} f_k = 1$.

Let the u_j's, v_k's, ϕ_{ij}'s, α_{jk}^i's, f and w be defined as before. Also let $\gamma_{jk}^i = I(L_i < u_j \le R_i \, , \, B_i^1 < u_j + v_k \le B_i^2)$, $j = 1, ..., r$, $k = 1, ..., s$. Then the conditional likelihood function of the observed data given $X_i \,\epsilon\, (L_i, R_i]$ has the form

$$L_C(w, f) = \prod_{i=1}^n \frac{\sum_{j=1}^r \sum_{k=1}^s \alpha_{jk}^i \, w_j \, f_k}{\sum_{j=1}^r \sum_{k=1}^s \gamma_{jk}^i \, w_j \, f_k} \, .$$

Define $\alpha_{ik} = I(v_k \in (U_i - R_i, V_i - L_i])$, $\beta_{ik} = I(v_k \in (B_i^1 - R_i, B_i^2 - L_i])$,

$$\phi_{ij}^* = \begin{cases} \sum_{k: U_i - u_j < v_k \le V_i - u_j} f_k & if \;\; u_j \,\epsilon\, (L_i \, , \, R_i] \\ 1 & otherwise \end{cases}$$

and

$$\eta_{ij}^* = \begin{cases} \sum_{k: B_i^1 - u_j < v_k \le B_i^2 - u_j} f_k & if \;\; u_j \,\epsilon\, (L_i, R_i] \\ 1 & otherwise \, , \end{cases}$$

$j = 1, \ldots, r$, $i = 1, \ldots, n$. Also define

$$\alpha_{ik}^* = \begin{cases} \sum_{U_i < u_l + v_k \le V_i} w_l & if \;\; v_k \,\epsilon\, (U_i - R_i \, , \, V_i - L_i] \\ 1 & otherwise \end{cases}$$

and

$$\beta_{ik}^* = \begin{cases} \sum_{B_i^1 < u_l + v_k \le B_i^2} w_l & if \;\; v_k \,\epsilon\, (B_i^1 - R_i, B_i^2 - L_i] \\ 1 & otherwise \, , \end{cases}$$

$k = 1, ..., s$, $i = 1, \ldots, n$. Then L_C can be rewritten as

$$L_C = \prod_{i=1}^n \frac{\sum_{j=1}^r \phi_{ij} \, \phi_{ij}^* \, w_j}{\sum_{j=1}^r \phi_{ij} \, \eta_{ij}^* \, w_j} = \prod_{i=1}^n \frac{\sum_{k=1}^s \alpha_{ik} \, \alpha_{ik}^* \, f_k}{\sum_{j=1}^s \beta_{ik} \, \beta_{ik}^* \, f_k}$$

with respect to w and f, respectively. For estimation of w and f, motivated by the self-consistency algorithm given in Section 3.4 and the maximum likelihood approach given above, Sun (1997b) proposes the following two-step self-consistency procedure.

Step 0. Estimate w as in step 1 of the two-step procedure given in Section 8.2.2 and denote the estimate by $w^{(0)} = \{ w_j^{(0)} \}$.
Step 1. At the lth iteration, define the updated estimate denoted by $f^{(l)} = \{ f_k^{(l)} \}$ of f as the maximum likelihood estimator of f from the conditional likelihood function L_C with assuming that w is known and $w_j = w_j^{(l-1)}$. It can be obtained by iterating the following self-consistency equation

$$f_k^{(b)} = \frac{1}{M(f^{(b-1)}, w^{(l-1)})} \sum_{i=1}^{n} [\mu_{ik}(f^{(b-1)}, w^{(l-1)}) + \nu_{ik}(f^{(b-1)}, w^{(l-1)})] \,,$$

$k = 1, ..., s$, with respect to b until convergence and with the $f_k^{(l-1)}$'s as the initial estimators. Here

$$\mu_{ik}(f, w) = \frac{\alpha_{ik}\, \alpha_{ik}^*\, f_k}{\sum_{j=1}^{s} \alpha_{ij}\, \alpha_{ij}^*\, f_j} \,, \quad \nu_{ik}(f, w) = \frac{(1 - \beta_{ik}\, \beta_{ik}^*)\, f_k}{\sum_{j=1}^{s} \beta_{ij}\, \beta_{ij}^*\, f_j}$$

and $M(f, w) = \sum_{i=1}^{n} \sum_{k=1}^{s} [\mu_{ik}(f, w) + \nu_{ik}(f, w)]$.
Step 2. Define the updated estimator denoted by $w^{(l)} = \{ w_j^{(l)} \}$ of w as the maximum likelihood estimator of w from L_C with assuming that f is known and $f_k = f_k^{(l)}$ from step 1. It can be obtained by iterating the following self-consistency equation,

$$w_j^{(b)} = \frac{1}{M^*(w^{(b-1)}, f^{(l)})} \sum_{i=1}^{n} [\mu_{ij}^*(w^{(b-1)}, f^{(l)}) + \nu_{ij}^*(w^{(b-1)}, f^{(l)})] \,,$$

$j = 1, ..., r$, with respect to b until convergence and with $w^{(l-1)}$ as the initial estimators. In the above,

$$\mu_{ij}^*(w, f) = \frac{\phi_{ij}\, \phi_{ij}^*\, w_j}{\sum_{k=1}^{r} \phi_{ik}\, \phi_{ik}^*\, w_k} \,, \quad \nu_{ij}^*(w, f) = \frac{(1 - \phi_{ij}\, \eta_{ij}^*)\, w_j}{\sum_{k=1}^{r} \phi_{ik}\, \eta_{ik}^*\, w_k}$$

and $M^*(w, f) = \sum_{i=1}^{n} \sum_{j=1}^{r} [\mu_{ij}^*(w, f) + \nu_{ij}^*(w, f)]$.
Step 3. Repeat steps 1 and 2 until the desired convergence occurs.

Sun (1997b) shows that the conditional likelihood function L_C increases at each iteration between steps 1 and 2 and thus the algorithm converges to a local or maximum of L_C. Note that for the case of no truncation, the conditional likelihood L_C differs from the full likelihood function L_F and it can seen from L_C that this may affect estimation of w, but not estimation of f. Numerical results support this. Also note that like the two-step approach, the conditional likelihood-based approach given here also operates through interval-censored data rather than doubly censored data. But the former method involves just one iteration, while the latter approach involves multiple iterations.

For the situation considered here, an alternative to the conditional likelihood-based approach is to maximize the full likelihood function

$$\prod_{i=1}^{n} \frac{\sum_{j=1}^{r} \sum_{k=1}^{s} \alpha_{jk}^{i} \, w_j \, f_k}{\sum_{j=1}^{j_{i0}} \sum_{k=k_{ij1}}^{k_{ij2}} \, w_j \, f_k} \, .$$

Here $j_{i0} \, = \, \max\{ \, j \, ; \, u_j \, \leq \, B_i^2 \, \}$, $k_{ij1} \, = \, \min\{ \, k \, ; \, v_k \, \geq \, \max\{ \, 0 \, , \, B_i^1 \, - \, u_j \, \} \, \}$, and $k_{ij2} \, = \, \max\{ \, k \, ; \, v_k \, \leq \, B_i^2 \, - \, u_j \, \}$. A major difference between L_C and the this full likelihood is that if the X_i's are observed exactly, as expected and one would like to, L_C is independent of w. That is, in this case, the conditional likelihood-based approach deals only with f, the parameters of interest. In contrast, the estimation procedure based on the latter still has to deal with both f and w. Tu (1995) investigated this latter approach for the case when all truncation intervals are identical.

Note that the three algorithms given above also apply to situations where the underlying X_i, S_i and T_i are continuous variables. This is because for a finite sample, the likelihood function has the same form for both discrete or continuous survival variables.

8.3 Semiparametric Regression Analysis

This section considers regression analysis of doubly censored failure time data. Inference approaches are discussed for regression parameters in two commonly used regression models, the PH and additive hazards models. For the former, we deal with both discrete and continuous versions of the model and different inference methods are described.

Let X_i, S_i, and T_i be defined as in the previous section and suppose that observed data from n independent subjects have the form

$$\{ \, (L_i, R_i] \, , \, (U_i, V_i] \, , \, \boldsymbol{Z}_i \, , \, i \, = \, 1, ..., n \, \} \, .$$

Here $(L_i, R_i]$ and $(U_i, V_i]$ denote observed intervals for X_i and S_i, respectively, and $\boldsymbol{Z}_i$ is a vector of covariates associated with subject i. For estimation of covariate effects on the failure time T_i's, we first consider the discrete PH model for the situation where all of X_i, S_i, and T_i are discrete survival variables. Estimation procedures are then discussed for the situations where X_i, S_i, and T_i are continuous and the covariate effect can be described by the PH model or the additive hazards model.

8.3.1 Analysis with the Discrete Proportional Hazards Model

In this subsection, we suppose that for subjects i, the effect of covariates on T_i can be described by model (1.10). That is, we have

$$S_k(\boldsymbol{Z}_i) \, = \, Pr(T_i > v_k | \boldsymbol{Z}_i) \, = \, (q_1 \cdots q_k)^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} \, , \; k \, = \, 1, ..., s - 1$$

given $\boldsymbol{Z}_i$. In this model, $v_1 \, < \, ... \, < \, v_s$ denote the possible values of the T_i's as before, $\boldsymbol{\beta}$ is the vector of regression parameters and

$$q_k \;=\; Pr(T_i > v_k | T_i > v_{k-1}, \boldsymbol{Z}_i = 0)$$

with $v_0 \;=\; 0$. Using the notation defined in the previous section, one can write the full likelihood function as

$$L_F(w, \boldsymbol{q}, \boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} \sum_{j=1}^{r} \sum_{k=1}^{s} \alpha_{jk}^{i}\, w_j\, f_k(\boldsymbol{Z}_i)\,, \tag{8.5}$$

where $\boldsymbol{q} \;=\; \{q_1, ..., q_{s-1}\}$ and

$$f_k(\boldsymbol{Z}_i) \;=\; (q_1 \cdots q_{k-1})^{\exp(\boldsymbol{Z}'_i \boldsymbol{\beta})}\, [1 \;-\; q_k^{\exp(\boldsymbol{Z}'_i \boldsymbol{\beta})}]$$

with $q_0 \;=\; 1$ and $q_s \;=\; 0$.

To estimate the parameters w_j's, q_k's and $\boldsymbol{\beta}$, one can use the maximum likelihood approach. For this, as in Section 1.4.8 and 7.4.1, it is common to reparameterize q_k by using, for example, $\gamma_k \;=\; \log[-\,\log(q_k)]$ to remove the range restriction on the q_k's and to improve convergence in the estimation process. Using the new parameters γ_k's, one has

$$f_k(\boldsymbol{Z}_i) \;=\; e^{-\,(e^{\gamma_1} + \ldots + e^{\gamma_{k-1}})\exp(\boldsymbol{Z}'_i \boldsymbol{\beta})}\, (1 \;-\; e^{-\exp(\gamma_k + \boldsymbol{Z}'_i \boldsymbol{\beta})})\;.$$

Another common reparameterization is to let $\gamma_k \;=\; \log[-\,\log(q_1 ... q_k)]$ and under this, one has

$$f_k(\boldsymbol{Z}_i) \;=\; e^{-\exp(\gamma_{k-1} + \boldsymbol{Z}'_i \boldsymbol{\beta})} \;-\; e^{-\exp(\gamma_k + \boldsymbol{Z}'_i \boldsymbol{\beta})}\;.$$

Let $\gamma \;=\; \{\,\gamma_k\,\}$. To maximize L_F given in (8.5) with respect to w, γ and $\boldsymbol{\beta}$, one can use the Newton-Raphson algorithm to solve the score functions from L_F. To reduce the dimension of the parameters, alternatively, one could use the following iterative two-step procedure.

Step 1. Choose initial estimators of $(w, \gamma, \boldsymbol{\beta})$.

Step 2. At the lth iteration, fix γ and $\boldsymbol{\beta}$ and let $\gamma_k \;=\; \hat{\gamma}_k^{(l-1)}$ and $\boldsymbol{\beta} \;=\; \hat{\boldsymbol{\beta}}^{(l-1)}$ from the previous iteration. Then define the updated estimator of w as the maximum likelihood estimator from $L_F(w, \gamma, \boldsymbol{\beta})$.

Step 3. Fix w and set $w_j \;=\; \hat{w}_j^{(l)}$ from step 2. Then define the updated estimators of γ and $\boldsymbol{\beta}$ as the maximum likelihood estimators from $L_F(w, \gamma, \boldsymbol{\beta})$.

Step 4. Repeat steps 2 and 3 until the desired convergence occurs.

The idea behind this algorithm was originally given in Kim et al. (1993), and as in Section 8.2.1, the resulting estimator could be a saddle point. Also as in Section 8.2.1, one can identify them by examining the signs of the eigenvalues of the information matrix. Given the estimators of $(w'_j s, \gamma'_k s, \boldsymbol{\beta})$, one could estimate their covariance matrix by the inverse of the observed Fisher information matrix. For the selection of initial estimators and the convergence criterion, one can use the approaches discussed in Section 8.2.1.

To implement step 2 of the two-step procedure above, we can use the self-consistency algorithm given in Section 3.4. To see this, note that one can rewrite the full likelihood function in (8.5) as

$$L_F(w, \gamma, \boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{r} \phi_{ij} \, \phi_{ij}^*(\gamma, \boldsymbol{\beta}) \, w_j \, ,$$

where $\phi_{ij} = I(u_j \in (L_i, R_i])$ and

$$\phi_{ij}^*(\gamma, \boldsymbol{\beta}) = \begin{cases} \sum_{k: U_i - u_j < v_k \leq V_i - u_j} f_k(\boldsymbol{Z}_i) & if \;\; u_j \, \epsilon \, (L_i \, , \, R_i] \\ 1 & otherwise \, . \end{cases}$$

That is, given γ and $\boldsymbol{\beta}$, $L_F(w, \gamma, \boldsymbol{\beta})$ can be regarded as a likelihood function from a set of interval-censored data. Hence the self-consistency algorithm given in Section 3.4 can be applied to estimate w using the estimator $\hat{w}^{(l-1)}$ from the previous iteration as the initial values.

For step 3, note that one also can rewrite $L_F(w, \gamma, \boldsymbol{\beta})$ as

$$L_F(w, \gamma, \boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{k=1}^{s} \alpha_{ik} \, \alpha_{ik}^* \, f_k(\boldsymbol{Z}_i) \, ,$$

where α_{ik} and α_{ik}^* are defined as in Section 8.2.3. This has the same form as the full likelihood function given in Section 6.2. Thus as in Section 6.2, one can apply the Newton-Raphson algorithm to maximize $L_F(w, \gamma, \boldsymbol{\beta})$. In doing this, the estimators of γ and $\boldsymbol{\beta}$ from the previous iteration can be used as initial values because w is fixed. Furthermore, for the first and second derivatives of the log likelihood function $l_F(w, \gamma, \boldsymbol{\beta}) = \log L_F(w, \gamma, \boldsymbol{\beta})$, the formulas derived in Section 6.2 can be used directly with treating the $\alpha_{ik} \, \alpha_{ik}^*$'s as the α_{ik}'s over there. In particular, for the case of $\boldsymbol{\beta} = 0$, one has the self-consistency equation

$$\hat{S}_0(0) = 1 \;\; , \;\; \hat{S}_k(0) = \hat{q}_k \, \hat{S}_{k-1}(\boldsymbol{Z}_i = 0) \;\; , \;\; \hat{q}_k = \frac{n_k' - d_k'}{n_k'}$$

for the estimator of the baseline survival function $S_k(\boldsymbol{Z}_i = 0)$, $k = 1, ..., s-1$. Here

$$d_k' = \sum_{i=1}^{n} \frac{\alpha_{ik} \, \hat{\alpha}_{ik}^* \, \hat{f}_k(0)}{\sum_{m=1}^{s} \alpha_{im} \, \hat{\alpha}_{im}^* \, \hat{f}_m(0)}$$

and

$$n_k' = \sum_{j=k}^{s} \sum_{i=1}^{n} \frac{\alpha_{ij} \, \hat{\alpha}_{ij}^* \, \hat{f}_j(0)}{\sum_{m=1}^{s} \alpha_{im} \, \hat{\alpha}_{im}^* \, \hat{f}_m(0)} \, ,$$

where $\hat{\alpha}_{ij}^*$ denotes α_{ij}^* with the w_j's replaced by their estimators. Note that here $\hat{f}_k(0) = \hat{S}_{k-1}(0) - \hat{S}_k(0)$.

An important application of this methodology is that one can derive a score test for the comparison of several survival functions. Suppose that study subjects come from $p + 1$ different populations, and let $\boldsymbol{Z}_i$ be the vector of group indicators. To compare the survival functions corresponding with the

$p + 1$ populations when only doubly censored data are available, one can use the score test statistic

$$\boldsymbol{U}_{st} \;=\; \frac{\partial l_F(w, \gamma, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{s} \frac{\alpha_{ij}\, \alpha_{ij}^{*}\, \boldsymbol{Z}_i\, [S_{k-1}(\boldsymbol{Z}_i) \log S_{k-1}(\boldsymbol{Z}_i) \;-\; S_k(\boldsymbol{Z}_i) \log S_k(\boldsymbol{Z}_i)]}{\sum_{k=1}^{s} \alpha_{ik}\, \alpha_{ik}^{*} \hat{f}_k(\boldsymbol{Z}_i)}$$

evaluated at $\boldsymbol{\beta} \,=\, 0$ and with w and the $S_k(0)$'s replaced by their estimators. A covariance estimator of $\boldsymbol{U}_{st}$ can be obtained similarly as in Section 6.2 from the observed Fisher information matrix.

8.3.2 Analysis with the Continuous Proportional Hazards Model

Suppose that the T_i's are continuous random variables and the hazard of T_i at time t is given by

$$\lambda_i(t) \;=\; \lambda_0(t)\, \exp(\, \boldsymbol{Z}_i' \boldsymbol{\beta}\,)$$

given $\boldsymbol{Z}_i$. Here as before, $\lambda_0(t)$ is an unknown baseline hazard function, and $\boldsymbol{\beta}$ denotes the vector of regression coefficients. For simplicity, we assume that the observations on the S_i's are right-censored and given by $S_i^* \,=\, \min\{\, S_i \,, C_i\,\}$ and $\delta_i \,=\, I(S_i \,=\, S_i^*)$, $i \,=\, 1, ..., n$. Here C_i is the censoring time associated with subject i and is assumed to be independent of S_i.

For estimation of β, define $Y_i(t \,|\, X_i) \,=\, I(\, S_i^* - X_i \,\geq\, t)$ and $N_i(t \,|\, X_i) \,=\, I(S_i^* - X_i \,\leq\, t\,,\, \delta_i = 1\,)$. Let $\boldsymbol{X} \,=\, (X_1, ..., X_n)$ and

$$S^{(j)}(t; \boldsymbol{\beta} \,|\, \boldsymbol{X}) \;=\; \frac{1}{n} \sum_{i=1}^{n} Y_i(t \,|\, X_i)\, \boldsymbol{Z}_i^j\, e^{\boldsymbol{Z}_i' \boldsymbol{\beta}} \;,$$

$j = 0, 1$, where $\boldsymbol{Z}_i^0 \,=\, 1$ and $\boldsymbol{Z}_i^1 \,=\, \boldsymbol{Z}_i$. Also let $\hat{H}$ denote the NPMLE of the cumulative distribution function of the X_i's based on interval-censored data on the X_i's only.

Note that if $\boldsymbol{X} \,=\, \boldsymbol{x} \,=\, (x_1, ..., x_n)$ is observed exactly, then one has right-censored data for the T_i's. Hence it is common to estimate $\boldsymbol{\beta}$ by the maximum partial likelihood estimator defined as the solution to $\boldsymbol{U}_p(\boldsymbol{\beta} \,|\, \boldsymbol{x}) \,=\, 0$, where

$$\boldsymbol{U}_p(\boldsymbol{\beta} \,|\, \boldsymbol{X}) \;=\; \int_0^{\tau} \sum_{i=1}^{n} \left[Z_i \;-\; \frac{S^{(1)}(t; \boldsymbol{\beta} \,|\, \boldsymbol{X})}{S^{(0)}(t; \boldsymbol{\beta} \,|\, \boldsymbol{X})} \right] d\, N_i(t \,|\, X_i) \qquad (8.6)$$

is the partial score function of β with τ denoting the longest possible follow-up time. Of course, $\boldsymbol{X}$ is unknown. By treating $\boldsymbol{X}$ as unknown parameters and using the profile likelihood idea, it is natural to estimate $\boldsymbol{\beta}$ by the solution, say $\hat{\boldsymbol{\beta}}_p$, to the following estimating equation

$$\boldsymbol{U}_p(\boldsymbol{\beta}, \hat{H}) \;=\; \left(\prod_{i=1}^{n} a_i^{-1} \right) \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} U_p(\boldsymbol{\beta} \,|\, \boldsymbol{x}) \prod_{i=1}^{n} \left[d\hat{H}(x_i) \right] \;=\; 0 \,, \quad (8.7)$$

where $a_i = \int_{L_i}^{R_i} d\hat{H}(x_i)$, $i = 1, ..., n$. It is easy to see that if the X_i's are observed exactly, $\hat{\boldsymbol{\beta}}_p$ reduces to the maximum partial likelihood estimator.

This estimation procedure was proposed in Sun et al. (1999), who proved that $\hat{\boldsymbol{\beta}}_p$ is consistent. Furthermore, $n^{1/2}\,(\,\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_0\,)$ has the asymptotic normal distribution with mean zero and covariance matrix that can be consistently estimated by $\hat{\Sigma}_p = A_p(\hat{\boldsymbol{\beta}}_p)\,\Gamma_p(\hat{\boldsymbol{\beta}}_p)\,A_p'(\hat{\boldsymbol{\beta}}_p)$. Here $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$, $A_p(\boldsymbol{\beta}) = \{-\,n^{-1}\,\partial U_p(\boldsymbol{\beta}, \hat{H})\,/\partial\boldsymbol{\beta}\,\}^{-1}$ and $\Gamma_p(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{b}}_{p\,i}(\boldsymbol{\beta})\,\hat{\boldsymbol{b}}_{p\,i}'(\boldsymbol{\beta})$, where

$$\hat{\boldsymbol{b}}_{p\,i}(\boldsymbol{\beta}) = \int_0^{\tau_1}\int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \left[\boldsymbol{Z}_i - \frac{S^{(1)}(t;\boldsymbol{\beta}\,|\,\boldsymbol{x})}{S^{(0)}(t;\boldsymbol{\beta}\,|\,\boldsymbol{x})}\right]\,[\,dN_i(t|x_i)$$

$$-\,\frac{Y_i(t|x_i)\exp(\,\boldsymbol{Z}_i'\boldsymbol{\beta})d\bar{N}(t|\boldsymbol{x})}{n\,S^{(0)}(t;\boldsymbol{\beta}\,|\,\boldsymbol{x})}\Bigg]\prod_{k=1}^n \frac{d\,\hat{H}(x_k)}{a_k}$$

and $\bar{N}(t|\boldsymbol{x}) = \sum_{i=1}^n N_i(t|x_i)$.

To determine $\hat{\boldsymbol{\beta}}_p$, one needs to solve the equation (8.7), and this could be difficult in practice. Alternatively, one can approximate or replace equation (8.7) by

$$\frac{1}{M}\sum_{l=1}^M \boldsymbol{U}_p(\boldsymbol{\beta}\,|\,\boldsymbol{x}_{(l)}) = 0\,,$$

where M is an integer and $\boldsymbol{x}_{(1)}, ..., \boldsymbol{x}_{(M)}$ are M sets of independent samples of $\boldsymbol{X}$ from $\hat{H}$ given the observed data. Similar to that discussed in Section 2.3, another alternative due to Pan (2001) is to use the multiple imputation approach as below.

Let M be an integer as before and for each l $(1 \le l \le M)$, let $X_i = x_{l\,i}$ be a random number from $\hat{H}$ given $X_i \,\epsilon\, (L_i, R_i]$. Then define $\hat{\boldsymbol{\beta}}_p^{(l)}$ as the solution to equation $\boldsymbol{U}_p(\boldsymbol{\beta}\,|\,\boldsymbol{x}) = 0$ and estimate $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}}_p = \frac{1}{M}\sum_{l=1}^M \hat{\boldsymbol{\beta}}_p^{(l)}$$

and the covariance matrix of $\tilde{\boldsymbol{\beta}}_p$ by

$$\tilde{\Sigma}_p = \frac{\sum_{l=1}^M I_{(l)}^{-1}}{M} + \left(1 + \frac{1}{M}\right)\frac{\sum_{l=1}^M (\hat{\boldsymbol{\beta}}_p^{(l)} - \tilde{\boldsymbol{\beta}}_p)\,(\hat{\boldsymbol{\beta}}_p^{(l)} - \tilde{\boldsymbol{\beta}}_p)'}{M\,-\,1}\,,$$

where $\boldsymbol{U}_p(\boldsymbol{\beta}\,|\,\boldsymbol{x})$ is defined in (8.6) and

$$I_{(l)}^{-1} = \left[-\,\frac{\partial U_p(\boldsymbol{\beta}|\boldsymbol{X}_{(l)})}{\partial\boldsymbol{\beta}'}\,\Bigg|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}_p}\right]^{-1}.$$

The estimator $\tilde{\boldsymbol{\beta}}_p$ is the average of M imputed estimators, and the first and second terms of its covariance estimator represent the estimates of, respectively, the within-imputation covariance and the between-imputation covariance. Pan (2001) suggests that the inference about $\boldsymbol{\beta}$ can be performed by using the normal approximation $N(\,\boldsymbol{\beta}\,,\,\tilde{\Sigma}_p\,)$ for $\tilde{\boldsymbol{\beta}}_p$.

As with the method discussed in Section 2.3, this Monte Carlo imputation approach reduces the interval-censored data problem to a right-censored data problem. Thus one can take advantage of existing software for right-censored data. In other words, both $\tilde{\boldsymbol{\beta}}_p$ and its covariance matrix estimator $\tilde{\Sigma}_p$ can be easily obtained compared to $\hat{\boldsymbol{\beta}}_p$ and $\hat{\Sigma}_p$. However, they could be less accurate than $\hat{\boldsymbol{\beta}}_p$ and $\hat{\Sigma}_p$.

Sometimes it also is interesting to estimate the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$. Once given $\hat{\boldsymbol{\beta}}_p$, a natural estimator of it is given by

$$\hat{\Lambda}_{p\,0}(t) \;=\; \int_{L_1}^{R_1} \ldots \int_{L_n}^{R_n} \int_0^t \frac{\sum_{i=1}^n d\,N_i(s\,|\,x_i)}{n\,S^{(0)}(s;\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{x})} \prod_{i=1}^n a_i^{-1}\, d\,\hat{H}(x_i)\,,$$

or one can estimate $\Lambda_0(t)$ by

$$\tilde{\Lambda}_{p\,0}(t) \;=\; \frac{1}{M}\sum_{l=1}^M \int_0^t \frac{\sum_{i=1}^n d\,N_i(s\,|\,x_{l\,i})}{n\,S^{(0)}(s;\tilde{\boldsymbol{\beta}}_p^{(l)}\,|\,\boldsymbol{x}_{(l)})}$$

with the use of the Monte Carlo imputation approach. In the case where the X_i's are observed exactly, these two estimators reduce to the Nelson-Aalen estimator of $\Lambda_0(t)$.

8.3.3 Analysis with the Additive Hazards Model

Consider the same problem discussed in Section 8.3.2. But instead of the PH model, we assume that given $\boldsymbol{Z}_i$, the hazard of T_i at time t is given by the additive hazards model

$$\lambda(t) \;=\; \lambda_0(t) \,+\, \boldsymbol{Z}_i'\,\boldsymbol{\beta}\,, \tag{8.8}$$

where $\lambda_0(t)$ and $\boldsymbol{\beta}$ are defined as before. Let the $Y_i(t\,|\,X_i)$'s, $N_i(t\,|\,X_i)$'s and $\boldsymbol{X}$ be defined as in the previous subsection. Also let

$$\bar{\boldsymbol{Z}}(t\,|\,\boldsymbol{X}) \;=\; \frac{\sum_{i=1}^n Y_i(t\,|\,X_i)\,\boldsymbol{Z}_i}{\sum_{i=1}^n Y_i(t\,|\,X_i)}\;.$$

For estimation of $\boldsymbol{\beta}$, as in the case of the PH model, first consider the situation where $\boldsymbol{X}$ is observed exactly. In this case, a common estimator is given by the solution to the equation $\boldsymbol{U}_a(\boldsymbol{\beta}\,|\,\boldsymbol{X}) \,=\, 0$, where

$$\boldsymbol{U}_a(\boldsymbol{\beta}\,|\,\boldsymbol{X}) = \sum_{i=1}^n \int_0^\tau \left[\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t\,|\,\boldsymbol{X})\right]\,\left[dN_i(t\,|\,X_i) \,-\, Y_i(t\,|\,X_i)\boldsymbol{Z}_i'\boldsymbol{\beta}\,dt\right]\;,$$

which is derived using the martingale theory (Lin and Ying, 1994). Here unlike $\boldsymbol{U}_p(\boldsymbol{\beta} \,|\, \boldsymbol{X})$, $\boldsymbol{U}_a(\boldsymbol{\beta} \,|\, \boldsymbol{X})$ is not a partial likelihood score function.

By using $\boldsymbol{U}_a(\boldsymbol{\beta} \,|\, X)$, as with $\boldsymbol{U}_p(\boldsymbol{\beta}, \hat{H})$ for the PH model, one can define the corresponding estimating function as

$$\boldsymbol{U}_a(\boldsymbol{\beta}, \hat{H}) \,=\, (\prod_{i=1}^{n} a_i^{-1}) \int_{L_1}^{R_1} ... \int_{L_n}^{R_n} U_a(\boldsymbol{\beta} \,|\, \boldsymbol{x}) \prod_{i=1}^{n} \left[d\hat{H}(x_i) \right]$$

for estimation of $\boldsymbol{\beta}$ under model (8.8) (Sun, Kim, and Sun, 2004). Let $\hat{\boldsymbol{\beta}}_a$ denote the solution to $\boldsymbol{U}_a(\boldsymbol{\beta}, \hat{H}) \,=\, 0$. Then one has

$$\hat{\boldsymbol{\beta}}_a \,=\, \left[\sum_{i=1}^{n} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \int_0^{\tau} Y_i(t|x_i) \left\{ \boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}}(t|\boldsymbol{x}) \right\}^{\otimes 2} dt \prod_{k=1}^{n} d\hat{H}(x_k)/a_k \right]^{-1}$$

$$\times \left[\sum_{i=1}^{n} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \int_0^{\tau} \left\{ \boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t|\boldsymbol{x}) \right\} dN_i(t|x_i) \prod_{k=1}^{n} d\hat{H}(x_k)/a_k \right] ,$$

where for a column vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} \,=\, \boldsymbol{a}\,\boldsymbol{a}'$. It is interesting to note that unlike $\hat{\boldsymbol{\beta}}_p$, the estimator $\hat{\boldsymbol{\beta}}_a$ has a closed form.

The consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_a$ are given in Sun, Kim, and Sun (2004). In particular, as $n \,\to\, \infty$, $n^{1/2}\,(\hat{\boldsymbol{\beta}}_a \,-\, \boldsymbol{\beta}_0)$ converges in distribution to a normal random vector with mean zero and covariance matrix that can be consistently estimated by $A_a^{-1}(\hat{\boldsymbol{\beta}}_a)\, B_a(\hat{\boldsymbol{\beta}}_a)\, A_a^{-1}(\hat{\boldsymbol{\beta}}_a)$, where

$$A_a(\boldsymbol{\beta}) \,=\, \frac{1}{n} \sum_{i=1}^{n} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \int_0^{\tau} Y_i(t \,|\, x_i) \left[\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}}(t|x) \right]^{\otimes 2} dt \prod_{k=1}^{n} d\hat{H}(x_k)/a_k \,,$$

$$B_a(\boldsymbol{\beta}) \,=\, \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{b}}_{a\,i}(\boldsymbol{\beta})\, \hat{\boldsymbol{b}}'_{a\,i}(\boldsymbol{\beta}) \,,$$

$$\hat{\boldsymbol{b}}_{a\,i}(\boldsymbol{\beta}) \,=\, \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \int_0^{\tau} \left[\boldsymbol{Z}_i \,-\, \bar{\boldsymbol{Z}}(t \,|\, \boldsymbol{x}) \right] d\hat{M}_i(t \,|\, \boldsymbol{x}, \boldsymbol{\beta}) \prod_{k=1}^{n} d\hat{H}(x_k)/a_k \,,$$

$$\hat{M}_i(t \,|, \boldsymbol{x}, \boldsymbol{\beta}) \,=\, N_i(t \,|\, x_i) \,-\, \int_0^t Y_i(s \,|\, x_i) \left[d\hat{\Lambda}_0(s \,|\, \boldsymbol{x}, \boldsymbol{\beta}) \,+\, \boldsymbol{Z}_i' \boldsymbol{\beta}\, ds \right]$$

and

$$\hat{\Lambda}_0(t \,|\, \boldsymbol{x}, \boldsymbol{\beta}) \,=\, \int_0^t \frac{\sum_{i=1}^{n} \left[dN_i(s \,|\, x_i) \,-\, Y_i(s \,|\, x_i) \boldsymbol{Z}_i' \boldsymbol{\beta}\, ds \right]}{\sum_{i=1}^{n} Y_i(s \,|\, x_i)} \,.$$

As with the PH model, in practice, the determination of $\hat{\boldsymbol{\beta}}_a$ may be complicated. To overcome this and make use of the existing software for right-censored data, one could also use the Monte Carlo imputation approach described in Section 8.3.2.

Specifically, as before, let M be an integer, and for each l $(1 \leq l \leq M)$, let $X_i = x_{l\,i}$ be a random number from $\hat{H}$ given $X_i \,\epsilon\, (L_i, R_i]$. Then for given $\boldsymbol{X} = \boldsymbol{x}_{(l)} = (x_{l\,1}, ..., x_{l\,n})$, solve the estimating equation $U_a(\boldsymbol{\beta} \,|\, \boldsymbol{x}_{(l)}) = 0$, which gives

$$\hat{\boldsymbol{\beta}}_a^{(l)} = \left[\sum_{i=1}^{n} \int_0^{\tau} Y_i(t \,|\, x_{l\,i}) \{ \boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t \,|\, \boldsymbol{x}_{(l)}) \}^{\otimes 2} \, dt \right]^{-1}$$

$$\times \left[\sum_{i=1}^{n} \int_0^{\tau} \{ \boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t \,|\, \boldsymbol{x}_{(l)}) \} \, dN_i(t \,|\, x_{l\,i}) \right] .$$

This gives an estimator of $\boldsymbol{\beta}$ as

$$\tilde{\boldsymbol{\beta}}_a = \frac{1}{M} \sum_{l=1}^{M} \hat{\boldsymbol{\beta}}_a^{(l)} ,$$

which is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_a$.

As $\tilde{\boldsymbol{\beta}}_p$, the covariance matrix of $\tilde{\boldsymbol{\beta}}_a$ can be estimated by

$$\tilde{\Sigma}_a = \frac{\sum_{l=1}^{M} \tilde{A}_l^{-1} \tilde{B}_l \tilde{A}_l^{-1}}{M} + (1 + \frac{1}{M}) \frac{\sum_{l=1}^{M} (\hat{\boldsymbol{\beta}}_a^{(l)} - \tilde{\boldsymbol{\beta}}_a)(\hat{\boldsymbol{\beta}}_a^{(l)} - \tilde{\boldsymbol{\beta}}_a)'}{M - 1} ,$$

the sum of the within-imputation and between-imputation covariance estimates, where

$$\tilde{A}_l = \sum_{i=1}^{n} \int_0^{\tau} Y_i(t \,|\, x_{l\,i}) \left[\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t \,|\, \boldsymbol{x}_{(l)}) \right]^{\otimes 2} \, dt$$

and

$$\tilde{B}_l =, \sum_{i=1}^{n} \int_0^{\tau} \left[\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}(t \,|\, \boldsymbol{x}_{(l)}) \right]^{\otimes 2} \, dN_i(t \,|\, x_{l\,i}) .$$

The term $\tilde{A}_l^{-1} \tilde{B}_l \tilde{A}_l^{-1}$ in this summation is the covariance estimator of $\hat{\boldsymbol{\beta}}_a^{(l)}$ based on the imputed right-censored data.

For the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s) \, ds$, given $\hat{\boldsymbol{\beta}}_a$, it can be estimated by

$$\hat{\Lambda}_{a\,0}(t) = \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \hat{\Lambda}_0(t \,|\, \boldsymbol{x}, \hat{\boldsymbol{\beta}}_a) \prod_{i=1}^{n} d\hat{H}(x_i) / \hat{a}_i .$$

With the Monte Carlo imputation approach, one can estimate $\Lambda_0(t)$ by

$$\tilde{\Lambda}_{a\,0}(t) = \frac{1}{M} \sum_{l=1}^{M} \hat{\Lambda}_0(t \,|\, \boldsymbol{x}_{(l)}, \hat{\boldsymbol{\beta}}_a^{(l)}) .$$

8.3.4 Discussion

There exist several differences among the approaches discussed in Sections 8.3.1 to 8.3.3. One major difference is that the maximum likelihood method given in Section 8.3.1 involves estimation of both distribution functions and regression parameters. In contrast, the estimating equation approaches described in Sections 8.3.2 and 8.3.3 deal only with the regression parameters and do not have to estimate the distribution function. The latter is usually regarded as nuisance parameters. Furthermore, the involvement of a large number of parameters in the maximum likelihood approach usually makes both computation and study of its properties difficult. Another difference is that in practice, the estimate of the covariance matrix using the Fisher information matrix in Section 8.3.1 may not give realistic and useful results, especially for large data sets. In contrast, the covariance estimates given in Sections 8.3.2 and 8.3.3 do not have the same problem in general. An advantage of the maximum likelihood approach is that it could yield more efficient estimators than the two estimating equation approaches.

For the models considered in Sections 8.3.2 and 8.3.3, one could develop an alternative method for inference by employing the maximum likelihood approach used in Section 8.3.1. Goggins et al. (1999a) discussed this approach for the continuous PH model and proposed a Monte Carlo EM algorithm. Compared with the two estimating equation approaches, this approach has the the differences similar to those between the approaches in Sections 8.3.1 to 8.3.3.

One can increase the efficiency of the methods given in Sections 8.3.2 and 8.3.3 or generalize them by adding a weight process. For example, for the PH model, one could use the same estimating equation $\boldsymbol{U}_p(\boldsymbol{\beta}, \hat{H}) = 0$ with replacing the function given in (8.6) by

$$\boldsymbol{U}_p(\boldsymbol{\beta} \,|\, \boldsymbol{X}, W_i' s) = \int_0^{\tau} \sum_{i=1}^{n} W_i(t) \left[\boldsymbol{Z}_i - \frac{S^{(1)}(\boldsymbol{\beta}, t \,|\, \boldsymbol{X})}{S^{(0)}(\boldsymbol{\beta}, t \,|\, \boldsymbol{X})} \right] d N_i(t \,|\, X_i) \, .$$

Here the $W_i(t)$'s are weight processes that may depend on observed data.

One practical issue related to the methods discussed in this section is model checking or the selection between the PH model and the additive hazards model. For this, of course, the prior knowledge about the possible underlying model for the subject matter considered is very important. From the statistical point of view, this is still an open question for interval-censored data. Of course, one could transfer the problem to the right-censored data problem by using the imputation approach and apply the existing model checking or selection techniques for right-censored data. Pan (2001) provided some discussion on this for the assessment of the PH model. Also some discussion is given in Section 8.5.2 through an example.

8.4 Nonparametric Comparison of Survival Functions

For comparison of several survival functions based on doubly censored data, one way is to use the score test given in Section 8.3.1. Alternatively, this section describes a nonparametric test procedure, which applies the idea discussed in Sun (2001b) and is a generalization of the generalized log-rank test given in Section 4.3.1 for interval-censored data.

Let the X_i's, S_i's, and T_i's be defined as before and suppose that only doubly censored data on the T_i's are available and they have the form

$$\{ \, (L_i, R_i] \, , \, (U_i, V_i] \, , \, i \, = \, 1, ..., n \, \} \, ,$$

where $X_i \in (L_i, R_i]$ and $S_i \in (U_i, V_i]$. Furthermore, suppose that all variables are discrete and each study subject randomly receives one of $p + 1$ different treatments or comes from one of $p + 1$ different populations. We assume that the distributions of the X_i's are same among all subjects. The goal is to test the hypothesis H_0 : the $p + 1$ survival functions corresponding to the different treatments or populations are identical.

To generalize the test procedure given in Section 4.3.1, we start by constructing estimates of the observed and expected numbers of failures as before. For this, as in Section 8.2, let $u_1 < ... < u_r$ denote the possible mass points for the X_i's and $v_1 < ... < v_s$ the possible mass points for the T_i's. Also let the $\hat{w}_j$'s and $\hat{f}_k$'s denote the joint maximum likelihood estimators of the probability functions of the X_i's and T_i's under H_0 given in Section 8.2. Define $\alpha_{ik} \, = \, I(\, v_k \, \epsilon \, (U_i - R_i, V_i - L_i])$ and

$$\hat{\alpha}^*_{ik} \, = \, \begin{cases} \sum_{u_j + v_k \epsilon (U_i, V_i]} \hat{w}_j & if \;\; v_k \, \epsilon \, (U_i - R_i \, , \, V_i - L_i] \\ 1 & otherwise \end{cases} \, ,$$

$i \, = \, 1, ..., n$, $k \, = \, 1, ..., s$. Also define $\delta_i \, = \, I(V_i \, < \, v_s)$ and $\rho_{ik} \, = \, I(\, \delta_i \, = \, 0 \, , \, U_i \geq v_k \,)$ as in Section 4.3.1, $i \, = \, 1, ..., n$, $k \, = \, 1, ..., s$.

Then as in Section 4.3.1, one can estimate the overall observed failure and risk numbers at time v_k by

$$d_k = \sum_{i=1}^{n} \delta_i \, \frac{\alpha_{ik} \, \hat{\alpha}^*_{ik} \, \hat{f}_k}{\sum_{m=1}^{s} \, \alpha_{im} \, \hat{\alpha}^*_{im} \, \hat{f}_m}$$

and

$$n_k = \sum_{j=k}^{s} \sum_{i=1}^{n} \delta_i \, \frac{\alpha_{ij} \, \hat{\alpha}^*_{ij} \, \hat{f}_j}{\sum_{m=1}^{s} \, \alpha_{im} \, \hat{\alpha}^*_{im} \, \hat{f}_m} \, + \, \sum_{i=1}^{n} \, \rho_{ik} \, ,$$

respectively, $k \, = \, 1, ..., s$. For each treatment group $l \, = \, 1, ..., p + 1$, the corresponding estimates at time v_k are

$$d_{k \, l} = \sum_{i}^{l} \, \delta_i \, \frac{\alpha_{ik} \, \hat{\alpha}^*_{ik} \, \hat{f}_k}{\sum_{m=1}^{s} \, \alpha_{im} \, \hat{\alpha}^*_{im} \, \hat{f}_m}$$

and

$$n_{k\,l} = \sum_{j=k}^{s} \sum_{i}^{l} \delta_i \, \frac{\alpha_{ij}\, \hat{\alpha}^*_{ij}\, \hat{f}_j}{\sum_{m=1}^{s} \alpha_{im}\, \hat{\alpha}^*_{im}\, \hat{f}_m} + \sum_{i}^{l} \rho_{ik} \, ,$$

respectively, where $\sum_i^l$ denotes the summation over all subjects with treatment l. If one has exact observations on the X_i's, these estimates give the corresponding estimators defined in Section 4.3.1.

Following the test statistic defined in Section 4.3.1, we can test H_0 using the statistic $\boldsymbol{U} = (U_1, ..., U_{p+1})'$, where

$$U_l = \sum_{k=1}^{s-1} \left(d_{kl} - \frac{n_{kl}\, d_k}{n_k} \right) .$$

The covariance matrix of $\boldsymbol{U}$ can be estimated similarly too as in Section 4.3.1. Specifically, let M be a prespecified integer. For each b $(1 \le b \le M)$,
Step 1. Let $\{X_i^{(b)}\, ; \, i = 1, ..., n\}$ be an independent sample of size n such that $X_i^{(b)}$ is drawn from the conditional probability function

$$Pr\{\, X_i^{(b)} = u_j \,\} = \frac{\hat{w}_j}{\sum_{m\,:\, L_i < u_m \le R_i} \hat{w}_m} \, , \; u_j \,\epsilon\, (L_i, R_i]$$

given $X_i \,\epsilon\, (L_i, R_i]$, $i = 1, ..., n$.
Step 2. For the given $X_i^{(b)}$'s, let $\{(T_i^{(b)}, \delta_i^{(b)})\, ; \, i = 1, ..., n\}$ be an independent right-censored survival sample of size n such that if $\delta_i = 0$, let $T_i^{(b)} = U_i - X_i^{(b)}$ and $\delta_i^{(b)} = 0$. Otherwise, let $T_i^{(b)}$ be a random number drawn from the conditional probability function

$$Pr\{\, T_i^{(b)} = v\} = \frac{\hat{f}_k}{\sum_{m\,:\, U_i - X_i^{(b)} < v_m \le V_i - X_i^{(b)}} \hat{f}_m}$$

over the v_k's that belong to $(U_i - X_i^{(b)}, V_i - X_i^{(b)}]$ and $\delta_i^{(b)} = 1$, $i = 1, ..., n$. Here $\delta_i^{(b)}$ is the censoring indicator.
Step 3. Given right-censored failure time data $\{(T_i^{(b)}, \delta_i^{(b)})\, ; \, i = 1, ..., n\}$, corresponding with the d_k, n_k, d_{kl}, and n_{kl}, first calculate the observed failure and risk numbers from all subjects and from subjects in each treatment group denoted by the $d_k^{(b)}$, $n_k^{(b)}$, $d_{kl}^{(b)}$, and $n_{kl}^{(b)}$, respectively. Then calculate the statistic $\boldsymbol{U}$ denoted by $\boldsymbol{U}^{(b)}$ with replacing d_k, n_k, d_{kl}, and n_{kl} by $d_k^{(b)}$, $n_k^{(b)}$, $d_{kl}^{(b)}$, and $n_{kl}^{(b)}$, respectively. Also calculate the estimate of the covariance matrix of $\boldsymbol{U}^{(b)}$ given by $\hat{\boldsymbol{V}}^{(b)} = \hat{\boldsymbol{V}}_1^{(b)} + ... + \hat{\boldsymbol{V}}_{s-1}^{(b)}$, where $\hat{\boldsymbol{V}}_k^{(b)}$ is a $(p+1) \times (p+1)$ matrix with elements

$$(\hat{V}_k^{(b)})_{l\,l} = \frac{n_{k\,l}^{(b)}\, (n_k^{(b)} - n_{k\,l}^{(b)})\, d_k^{(b)}\, (n_k^{(b)} - d_k^{(b)})}{(n_k^{(b)})^2\, (n_k^{(b)} - 1)} \, , \; l = 1, ..., p+1$$

and

$$(\hat{V}_k^{(b)})_{l_1 l_2} = -\frac{n_{k\,l_1}^{(b)}\, n_{k\,l_2}^{(b)}\, d_k^{(b)}\,(n_k^{(b)} - d_k^{(b)})}{(n_k^{(b)})^2\,(n_k^{(b)} - 1)}\,,\ l_1 \neq l_2 = 1, ..., p+1$$

$k = 1, ..., s-1$.
Step 4. Repeat the steps 1 to 3 for each $b = 1, ..., M$ and then estimate the covariance matrix of $\boldsymbol{U}$ by $\hat{\boldsymbol{V}} = \hat{\boldsymbol{V}}_1 + \hat{\boldsymbol{V}}_2$. Here

$$\hat{\boldsymbol{V}}_1 = \frac{1}{M}\sum_{b=1}^{M}\hat{\boldsymbol{V}}^{(b)}$$

and

$$\hat{\boldsymbol{V}}_2 = \left(1 + \frac{1}{M}\right)\frac{\sum_{b=1}^{M}[\boldsymbol{U}^{(b)} - \bar{\boldsymbol{U}}]\,[\boldsymbol{U}^{(b)} - \bar{\boldsymbol{U}}]^t}{M - 1}\,,$$

where $\bar{\boldsymbol{U}} = \sum_{b=1}^{M}\boldsymbol{U}^{(b)}/M$.

Once $\boldsymbol{U}$ and $\hat{\boldsymbol{V}}$ are obtained, the test of the hypothesis H_0 can be carried out using the statistic $U^* = \boldsymbol{U}'\,\hat{\boldsymbol{V}}^{-}\,\boldsymbol{U}$, whose distribution can be approximated by the χ^2 distribution with p degrees of freedom, where $\hat{\boldsymbol{V}}^{-}$ denotes a generalized inverse of $\hat{\boldsymbol{V}}$. An equivalent test is to use any p elements of $\boldsymbol{U}$ and the corresponding submatrix of $\hat{\boldsymbol{V}}$. Sun (2001b) gave a procedure that is similar to the one given above, but does not reduce to the log-rank test for right-censored data.

The test procedure given above is closely related to the score test given in Section 8.3.1. To see the relationship between them, define $\boldsymbol{Z}_i$ to be the $p \times 1$ vector of treatment indicators such that $\boldsymbol{Z}_i = 0$ for subjects in population $p+1$, and for subjects in population l $(1 \leq l \leq p)$, its lth element is equal to 1 and all other elements are equal to 0. Also define

$$c_{dik} = \frac{\alpha_{ik}\,\alpha_{ik}^*\,\hat{f}_k}{\sum_{m=1}^{s}\hat{\alpha}_{im}\,\hat{\alpha}_{im}^*\,\hat{f}_m}$$

and $c_{rik} = \sum_{m=k}^{r} c_{dim}$, $k = 1, ..., s-1$, $i = 1, ..., n$. Note that the c_{dik} and c_{rik} are the conditional probabilities of the events $T_i = v_k$ and $T_i \geq v_k$ given the observed data, respectively.

As in Section 6.5.2, let D_k denote the set of subjects who have a non-zero probability of failing at v_k and R_k the set of subjects who have a non-zero probability of being at risk at v_k^- given the observed data, $k = 1, ..., s-1$. Then similarly as in Section 6.5.2, using the notation defined here, one can rewrite the score test statistic given in Section 8.3.1 for H_0 as

$$\boldsymbol{U}_{st} = \sum_{i=1}^{n}\sum_{k=1}^{s-1}\left(\boldsymbol{Z}_i\,c_{rik}\,\log\hat{\lambda}_k - \frac{\boldsymbol{Z}_i\,c_{dik}\,\log\hat{\lambda}_k}{1-\hat{\lambda}_k}\right)$$

$$= \sum_{k=1}^{s-1} \frac{-\log \hat{\lambda}_k}{1-\hat{\lambda}_k} \left(\sum_{i \epsilon D_k} \boldsymbol{Z}_i \, c_{dik} \; - \; (1-\hat{\lambda}_k) \sum_{i \epsilon R_k} \boldsymbol{Z}_i \, c_{rik} \right) ,$$

where $\hat{\lambda}_k = \sum_{m=k+1}^{s} \hat{f}_m / \sum_{m=k}^{s} \hat{f}_m$. In contrast, asymptotically, the test statistic $\boldsymbol{U}$ can be rewritten as

$$\boldsymbol{U} = \sum_{k=1}^{s-1} \left(\sum_{i \epsilon D_k} \boldsymbol{Z}_i \, c_{dik} \; - \; (1-\hat{\lambda}_k) \sum_{i \epsilon R_k} \boldsymbol{Z}_i \, c_{rik} \right)$$

assuming that all censoring intervals have finite lengths. Note that if the lengths of intervals $(v_{k-1}, u_k]$ are very small as $n \to \infty$, then $\log \hat{\lambda}_k^{-1}$ will be approximately equal to $1-\hat{\lambda}_k$. Thus $\boldsymbol{U}$ and $\boldsymbol{U}_{st}$ are asymptotically equivalent.

8.5 Examples

This section presents two illustrative examples for the methods discussed in the previous sections. The first one concerns a set of doubly censored failure time data arising from an AIDS clinical trial with the focus on the duration of viral suppression. The second example considers the data discussed in Section 1.2.3 from an AIDS cohort study regarding AIDS latency time.

8.5.1 Analysis of Duration of Viral Suppression Data

Table 8.1 presents a set of doubly censored failure time data arising from the same AIDS clinical trial, ACTG 359, that generated the data given in data set III of Appendix A. The table gives the observed information about the duration of viral suppression defined as the period of time during which the number of RNA copies is below the threshold of 500 viral copies/ml. As the time at which the number of RNA copies first drops below the threshold discussed in Section 6.3.2, the duration of viral suppression is another variable of great interest to clinicians, which is often used to measure the effectiveness of AIDS treatments among other purposes. As discussed in Section 6.3.2, viral load is usually only measured periodically, and thus the times at which a person's viral load falls below and comes back over again 500 copies are not exactly observed. In other words, only doubly censored data are available for the duration of viral suppression as given in Table 8.1. Unlike in Section 6.3.2, here we analyze all patients together without considering their initial RNA levels and focus on the nonparametric estimation problem.

Let T denote the duration of viral suppression and X and S times at which an AIDS patient's RNA falls below and comes back over again 500 copies, respectively. Then $T = S - X$. Table 8.1 gives observed intervals $(L, R]$ and $(U, V]$ for X and S, respectively, of 124 AIDS patients from ACTG 359 whose numbers of RNA copies were measured at least once and fell below

Table 8.1. Observed intervals in months given by $(L, R]$ and $(U, V]$ for the duration of viral suppression of 124 AIDS patients from ACTG 359 with . indicating that RNA is still below 500 at month 12

L	R	U	V	L	R	U	V	L	R	U	V	L	R	U	V
1	2	2	3	1	2	2	3	0	1	1	2	0	1	1	2
8	10	12	.	0	1	6	8	0	1	1	2	0	1	4	6
0	1	1	2	0	2	8	12	0	1	2	.	1	2	2	3
0	1	2	3	0	1	12	.	0	1	2	4	2	3	12	.
1	2	2	3	1	2	3	4	0	1	1	2	0	1	12	.
0	1	12	.	0	1	12	.	1	2	12	.	0	1	1	2
2	3	4	6	2	3	12	.	1	2	2	3	0	1	12	.
1	2	10	12	2	3	4	6	0	1	1	2	2	3	12	.
2	3	3	4	1	2	12	.	0	1	12	.	0	1	4	6
0	1	1	2	0	1	2	3	1	4	4	6	0	1	12	.
0	1	2	3	0	1	12	.	0	1	8	10	0	1	1	2
3	4	4	8	0	1	10	12	8	10	12	.	2	3	3	4
1	2	12	.	0	1	12	.	0	1	1	2	0	2	12	.
1	2	12	.	1	2	3	4	3	4	12	.	1	3	6	.
1	2	12	.	0	2	6	8	1	2	4	6	0	1	4	6
0	2	3	4	3	4	6	8	0	1	1	2	1	2	12	.
0	1	1	2	1	2	8	.	0	2	2	4	0	1	12	.
1	2	12	.	1	2	2	3	0	1	1	2	2	3	12	.
0	1	1	2	0	1	1	2	0	1	1	2	1	2	2	3
4	6	12	.	0	1	3	4	0	1	2	3	0	1	4	6
0	1	12	.	0	1	1	2	1	2	2	3	2	3	4	.
0	1	12	.	2	3	10	.	1	2	2	3	3	4	4	6
6	8	10	.	1	2	10	12	0	1	3	4	2	3	6	.
0	1	6	8	0	1	3	4	3	4	4	6	6	8	12	.
1	2	2	4	2	3	3	6	0	1	4	6	1	2	12	.
1	3	3	4	0	4	4	6	0	1	4	6	1	2	3	4
1	2	4	6	0	1	3	4	0	1	12	.	0	1	1	2
0	2	12	.	2	3	4	8	1	2	6	8	0	1	12	.
0	1	12	.	0	1	8	10	0	1	4	6	0	1	1	2
0	1	12	.	2	3	6	8	2	3	6	8	0	1	12	.
0	1	3	4	3	4	4	6	0	1	2	3	1	2	6	8

500 during the 12 months period. As with data set III of Appendix A, the time unit is month. For example, the observation $(0, 1]$ and $(12, .]$ means that the patient's RNA falls below 500 at the very first clinical visit, the end of the first month, and is still below 500 at the last visit, the end of month 12. In other words, his or her duration of viral suppression is greater than 12 months. For the patient with observation $(1, 2]$ and $(2, 3]$, his or her RNA is above 500 at the first visit (the end of the first month), falls below 500 at the second visit (the end of the second month), and then jumps over 500 before the end of month 3.

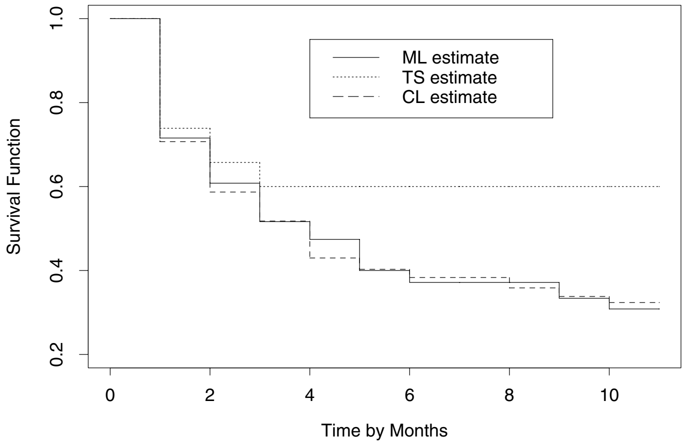

Fig. 8.1. Estimates of the survival function of the duration of viral suppression.

Figure 8.1 presents estimates of the survival function of the duration of viral suppression given by the maximum likelihood (ML), two-step (TS), and conditional likelihood-based (CL) approaches described in Section 8.2. For these estimates, we use the self-consistent estimates given in Section 3.4 based on separate interval-censored data as initial estimates. Other initial estimates such as uniform distributions are also studied and give similar results. The figure shows that the ML and CL estimates are quite close to each other, but the TS estimate seems to overestimate the duration of viral suppression. Both ML and CL estimates suggest that the median duration of viral suppression is about 4 months.

To explore the difference between ML and CL estimates and TS estimate shown in Figure 8.1, Figure 8.2 gives the resulting estimates of the survival function of the initial event, the RNA copies of the patients in the study falling below 500, from the three approaches. It is interesting to note that the ML and TS approaches give almost identical estimates. This indicates that the use of the marginal estimate of the survival function of the initial event in the TS approach is not the only difference between the TS and ML and CL approaches. As expected, the ML and CL approaches can yield different estimates of the survival function of the initial event.

8.5.2 Analysis of AIDS Latency Time Data

To analyze the doubly censored data discussed in Section 1.2.3, we first estimate the survival functions of AIDS latency times of the patients in the two treatment groups separately. Figure 8.3 displays the estimates obtained

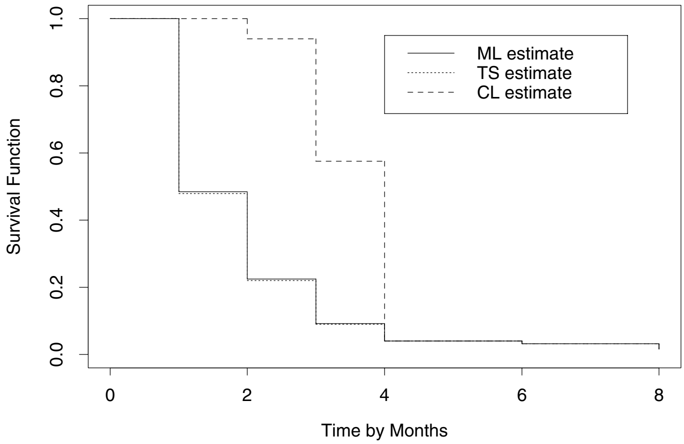

Fig. 8.2. Estimates of the survival function of the initial event for the duration of viral suppression.

using the three approaches described in Section 8.2 corresponding with the two treatment groups. It can be seen from the figure that all three methods give similar estimates for each of the two groups. All three estimates of the survival function for the patients in the lightly treated group are on top of these for the patients in the heavily treated group. That is, the patients in the lightly treated group have longer AIDS latency time than those in the heavily treated group. In other words, the HIV patients in the lightly treated group seem to enjoy longer life without AIDS than those in the heavily treated group, and the more contaminated blood factor means earlier onset of AIDS. For the comparison of the survival functions of HIV infection time between the two groups, Figure 8.4 gives their estimates obtained using the ML approach and indicates that the patients in the two groups seem to have similar HIV infection rates. For the estimates given in both Figures 8.3 and 8.4, as with those in Figures 8.1 and 8.2, the self-consistent estimates based on separate interval-censored data are used as initial estimates.

Now we consider the comparison of the survival functions of the AIDS latency times for patients in the two treatment groups. For this, with letting the lightly treated group being group 1, the application of the generalized log-rank test described in Section 8.4 gives $U_1 = -6.9884$ and $U^* = 4.8754$ with $M = 100$. Based on the χ^2 distribution with degree of freedom 1, these results correspond with a p-value of 0.027. They suggest that the patients in the two groups have significantly different survival functions of the AIDS latency time and confirm what is shown in Figure 8.3.

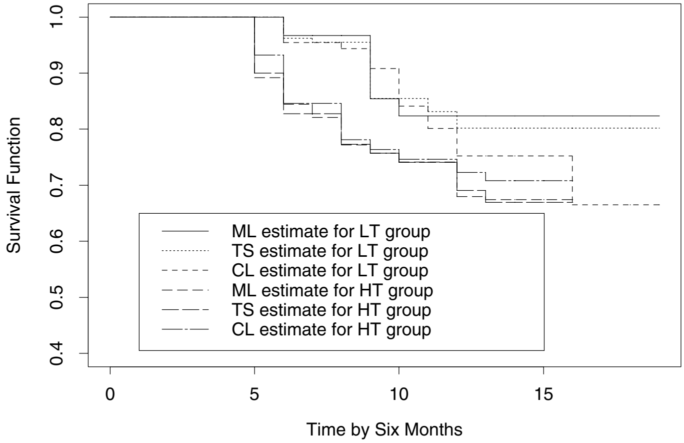

Fig. 8.3. Estimates of the survival functions of AIDS latency time.

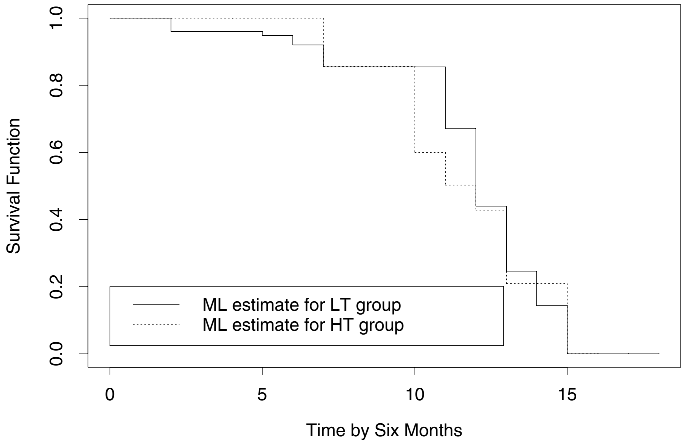

Fig. 8.4. Estimates of the survival functions of HIV infection time.

To estimate the effect of the amount of the blood factor on AIDS latency time, we apply the inference approaches described in Sections 8.3.2 and 8.3.3 to the data by assuming that the AIDS latency time follows the continuous PH or additive model. Define $Z_i = 0$ if subject i belongs to the lightly treated group and 1 otherwise, $i = 1, ..., 188$. Then using the continuous PH model with $M = 100$, we obtain $\tilde{\beta}_p = 0.7087$ with estimated standard error equal to 0.5083, yielding a p-value of 0.163 for testing $\beta = 0$. If using the additive hazards model with $M = 100$, we obtain $\tilde{\beta}_a = 0.0139$, and its estimated standard deviation is 0.0062, giving a p-value of 0.025 for testing $\beta = 0$. For both models, we tried larger values of M and got similar results.

Compared with the estimates given in Figure 8.3 and the result obtained using the generalized log-rank test, the result from the additive hazards model seems reasonable. In contrast, the result given by the continuous PH model

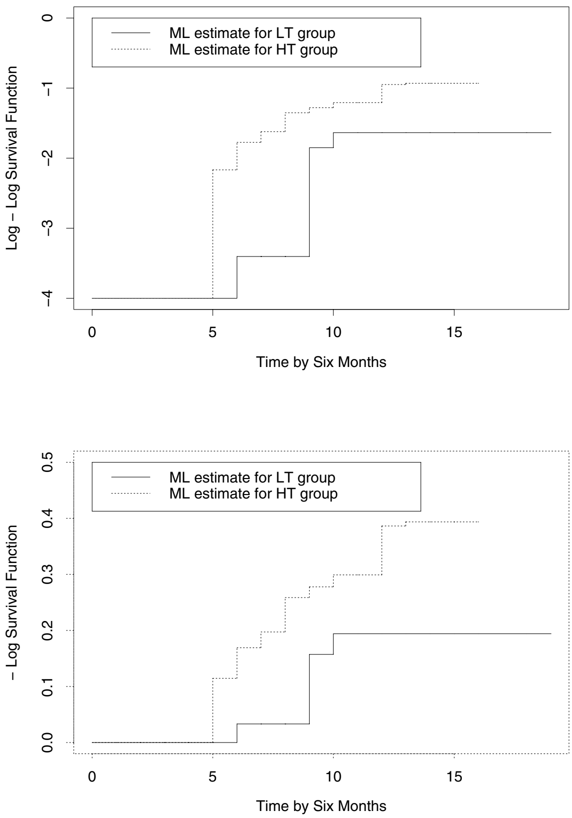

Fig. 8.5. Log - log and - log estimates of the survival functions of AIDS latency time.

seems to underestimate the effect of the blood factor on AIDS latency time. To understand this underestimation, instead of using the Monte Carlo imputation approach, we estimate the covariate effect directly from the equation (8.7) and obtain $\hat{\beta}_p = 0.7032$ with an estimated standard error 0.2834. It can be seen that the estimated effect of the blood factor is similar as before, but the estimated standard error is much smaller. This suggests that the Monte Carlo imputation approach may not affect estimation of covariate effects but could overestimate the variance. For the regression analysis here, of course, one also can use the maximum likelihood approaches discussed in Section 8.3, and they give similar results (Goggins et al., 1999a; Kim et al., 1993).

Note that for the two-sample situation, under the PH model, the log - log survival functions should be parallel to each other, while under the additive hazards model, the - log survival functions should be parallel to each other. To check which of the two models provides a better fit to the observed data here, Figure 8.5 presents both log - log and - log estimates of the two survival functions given by the ML approach in Figure 8.3 corresponding to the two treatment groups. From the figure, the additive hazards model may seem better, but the difference between the two models does not seem to be significant.

8.6 Bibliography, Discussion, and Remarks

Much of the research on doubly censored failure time data is motivated by AIDS research and, in particular, by the seminal paper De Gruttola and Lagakos (1989), which studied estimation of the distribution function of AIDS latency time. Following them, a number of other authors discussed the same problem and these include Bacchetti (1990), Fang and Sun (2001), Frydman (1992, 1995a, b), Gómez and Calle (1999), Jewell (1994), Jewell et al. (1994), Joly and Commenges (1999), Leung and Elashoff (1996), Lim et al. (2002), Sternberg and Satten (1999), and Sun (1995) in addition to Gómez and Lagakos (1994), Sun (1997b), and Tu (1995). In particular, Bacchetti (1990) and Joly and Commenges (1999) considered the use of the penalized likelihood approach for estimation of the hazard function of the survival time of interest, and Fang and Sun (2001) established the consistency of the two-step estimation procedure given in Section 8.2.2. Frydman (1992, 1995a, b) and Leung and Elashoff (1996) applied the three-state model approach to the estimation problem, and Lim et al. (2002) gave an estimation procedure for the situation where there is a change-point in the survival function of the survival time of interest.

Articles that discussed other issues about doubly censored data include Goggins (1999a), Kim et al. (1993), Pan (2001), Sun (2001b, 2004), Sun et al. (1999), Sun, Kim, and Sun (1999), Sun, Lim, and Zhao (2004), and Zhu and Sun (2006). Among these, Sun (2001b) considered nonparametric comparison of several distribution functions, and Sun (2004) provided a relatively complete review of the field. Also, Sun, Lim, and Zhao (2004) discussed the test of the independence assumption between X and T used in the preceding sections, and Zhu and Sun (2006) presented several methods for pointwise estimation of variances of estimated survival functions. Other references gave semiparametric approaches for regression analysis of doubly censored data.

For simplicity of discussion, several methods discussed in the previous sections assume that the infection times or X_i's follow the same distribution for all subjects in the study. They can be easily generalized to situations where this does not hold. Consider situations, for example, where the distribution of the T_i's depends on covariates through the continuous PH model. Suppose that the cumulative distribution function H of the X_i's is unknown and may depend on covariates through regression parameters θ. Let $\hat{H}(x\,;\,\hat{\theta})$ denote some consistent estimates of H and θ based on the observed interval-censored on the X_i's. Then the equation (8.7) can be generalized to

$$(\prod_{i=1}^{n} a_l^{-1})\int_{L_1}^{R_1} \dots \int_{L_n}^{R_n} U_p(\boldsymbol{\beta}\,|\,\boldsymbol{x}) \prod_{i=1}^{n} \left[d\hat{H}(x_l\,;\,\hat{\theta}) \right] = 0$$

for estimation of β.

As mentioned above, a basic assumption behind all methods discussed in this chapter is that the time between the infection and the onset of a disease is independent of the infection time. It makes the analysis of doubly censored data tractable, but may not be true in practice (Sun, Lim, and Zhao, 2004). To assess this assumption, one can apply the test procedure given in Sun, Lim, and Sun (2004) if the dependence of the T_i's on the X_i's can be described by a PH model. If the independence does not hold, one analysis approach is to use the three-state Markov model (Frydman, 1992, 1995a, b; Leung and Elashoff, 1996). In this model, the three states can be infection-free, infection, and onset of the disease, respectively, with the third state as an absorbing state.

There are several other issues for the analysis of doubly censored failure time data that were not treated at all or in detail in this chapter. One is variance or covariance estimation for some nonparametric and semiparametric estimation approaches. For the problem, a common suggestion is to use the observed Fisher information matrix, and it is well-known that this approach could give unrealistic results when there exist a large number of parameters. An alternative is to use the resampling methods discussed in Zhu and Sun (2006). However, these methods are only for pointwise variance estimation. Also, for both Fisher information matrix and resampling approaches, there does not exist theoretical justification. The asymptotic property of the proposed methods for doubly censored data is another issue for which there is not much discussion both in this chapter and in the literature except Fang and Sun (2001).

This page intentionally blank

9

Analysis of Panel Count Data

9.1 Introduction

All of preceding chapters discuss situations where the event of interest or failure can occur only once and the response variable of interest is time to the event. In this chapter, we consider situations where the event or failure can occur multiple times or repeatedly. In other words, study subjects could experience recurrences of the same event. The studies that result in this type of information are often referred to as event history studies, and the resulting data are referred to as event history data. In addition to medical studies, other application areas that frequently produce event history data include reliability studies and social sciences, and there exist several books on the analysis of such data such as Nelson (2003) and Vermunt (1997).

The studies that deal with recurrent events can be generally classified into two types. One is the studies that monitor study subjects continuously and thus give recurrent event data (Cook and Lawless, 2006), which record the times of all occurrences of events. The other is the studies in which study subjects are checked or observed only at discrete time points and thus they produce panel count data, which give only the numbers of occurrences of the events between observation times. The latter type of studies is a combination of the former type of studies and interval-censoring and exists because, for example, it may be too expensive, impossible, or not realistic to conduct continuous follow-ups. Examples of recurrent event data include occurrences of hospitalizations of intravenous drug users (Wang, et al., 2001), occurrences of the same infection such as recurrent pyogenic infections among inherited disorder patients (Lin et al., 2000), repeated occurrences of certain tumors, and warranty claims for a particular automobile (Kalbfleisch et al., 1991). These examples become examples of panel count data if the continuous observation scheme is changed to a discrete observation scheme. Some specific examples of panel count data are discussed below.

For the analysis of recurrent event data, a number of statistical methods have been proposed. In addition to those mentioned above, other references on recurrent event data include Chen et al. (2004), Cook and Lawless (1996), Cook et al. (1996), Lawless and Nadeau (1995), Lawless (2001), Lin, Wei, and Ying (1998), Pepe and Cai (1993), and Wang and Chen (2000). Also, Andersen et al. (1993) is an excellent book that provides a comprehensive coverage of counting process approaches for the analysis of recurrent event data, and Kalbfleisch and Prentice (2002) devote one chapter to recurrent event data.

To analyze recurrent event data, it is common and convenient to characterize the occurrences of recurrent events by counting processes and to model the intensity process of the counting process. Consider a study consisting of a single type of recurrent event and suppose that there is a p-dimensional covariate process $\boldsymbol{Z}(t)$. Let $N^*(t)$ denote the number of occurrences of the event over the interval $[0, t]$ and $\mathcal{F}_{t-}$ the σ-field generated by $\{\, N^*(s)\,,\, \boldsymbol{Z}(s) \,:\, 0 \leq s < t \}$. Then the intensity process $\lambda(t\,;\, \boldsymbol{Z})$, assuming that it exists and occurrence times are absolutely continuous, of $N^*(t)$ associated with $\mathcal{F}_{t-}$ can be defined as

$$\lambda(t\,;\, \boldsymbol{Z}) \,=\, \lim_{\delta \downarrow 0} \delta^{-1}\, E\,[\, d\,N^*(t) \,|\, \mathcal{F}_{t-}\,]\,,$$

where $d\,N^*(t) \,=\, N^*\{(t+\delta)-\} \,-\, N^*(t-)$, the increment of $N^*(t)$ over the small interval $[t, t+\delta)$.

Various models can be used for $\lambda(t\,;\, \boldsymbol{Z})$ and perhaps the most commonly used one is the Andersen-Gill intensity model,

$$\lambda(t\,;\, \boldsymbol{Z}) \,=\, \lambda_0(t)\, \exp\,[\, \boldsymbol{Z}'(t)\, \boldsymbol{\beta}\,]\,,$$

proposed in the seminal paper Andersen and Gill (1982). Here $\lambda_0(t)$ is an unspecified continuous function and $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters. This model assumes that the history of the whole recurrent process affects the recurrence of the event at time t only through time-varying covariates at t in this multiplicative fashion. One way to relax this assumption is to drop the conditioning on the event history and in place of $\lambda(t\,;\, \boldsymbol{Z})$, model the marginal intensity $d\,\mu(t\,;\, \boldsymbol{Z}) \,=\, E\,[\, d\,N^*(t) \,|\, \boldsymbol{Z}(s),\, 0 \leq s < t\,]$ by

$$d\,\mu(t\,;\, \boldsymbol{Z}) \,=\, \exp\,[\, \boldsymbol{Z}'(t)\, \boldsymbol{\beta}\,]\, d\,\mu_0(t) \tag{9.1}$$

(Lin et al., 2000; Pepe and Cai, 1993), where $\mu_0(t)$ is an unknown continuous function. If $\boldsymbol{Z}$ is fixed or external covariates, which implies that

$$E\,[\, d\,N^*(u) \,|\, \boldsymbol{Z}(s),\, 0 \leq s < u\,] \,=\, E\,[\, d\,N^*(u) \,|\, \boldsymbol{Z}(s),\, 0 \leq s < t\,]$$

for all $t \geq u$ (Kalbfleisch and Prentice, 2002), then we have

$$\mu(t\,;\, \boldsymbol{Z}) \,=\, E\,[\, N^*(t) \,|\, \boldsymbol{Z}(s),\, s \geq 0\,]\,. \tag{9.2}$$

That is, $\mu(t\,;\, \boldsymbol{Z})$ represents the mean function of the recurrent process $N^*(t)$. In particular, if $\boldsymbol{Z}$ is time-independent, model (9.1) gives

$$\mu(t\,;\,\boldsymbol{Z}) \,=\, \mu_0(t)\,\exp(\boldsymbol{Z}'\boldsymbol{\beta})\,. \tag{9.3}$$

Models (9.1) and (9.3) are often referred to as the proportional rates and means models, respectively.

In most studies, study subjects are followed for limited amounts of time and there exists a variable C representing the follow-up or censoring time. Define $N(t) \,=\, N^*(t \wedge C)$ and $Y(t) \,=\, I(t \leq C)$, where $a \wedge b \,=\, \min(a, b)$. For most of the methods developed for recurrent event data, C is assumed to be independent of the underlying recurrent process, meaning that

$$E\,[\,d\,N(t)|N(s),\,Y(s),\,\boldsymbol{Z}(s),\,0 \leq s < t\,] \,=\, Y(t)\,E\,[\,d\,N^*(t)\,|\,\mathcal{F}_{t-}\,]$$

or

$$E\,[\,d\,N(t)|Y(t),\,\boldsymbol{Z}(t)\,] \,=\, Y(t)\,E\,[\,dN^*(t)\,|\,\boldsymbol{Z}(t)\,] \tag{9.4}$$

for all $t \,\geq\, 0$ (Lin et al., 2000).

For the analysis of panel count data, it is more convenient to work directly on the mean function $\mu(t\,;\,\boldsymbol{Z})$ defined in (9.2) due to the incomplete nature of observed information. In this case, a natural and simple approach is to fit the data to parametric Poisson processes or mixed parametric Poisson processes. For example, Hinde (1982) and Breslow (1984) discussed regression analysis of Poisson count data, and Thall (1988, 1989) gave some regression approaches for mixed Poisson processes. Another parametric approach for the analysis of panel count data is to treat them as longitudinal count data and to use the generalized estimating equation approach (Diggle et al., 1994; Thall and Vail, 1990). In this chapter, we focus attention on nonparametric and semiparametric approaches that regard observed data as realizations of some underlying counting processes.

Section 9.2 deals with one-sample analysis of panel count data with the focus on nonparametric estimation of the mean function $\mu(t\,;\,\boldsymbol{Z})$. Two approaches are discussed along with two illustrative examples. In Section 9.3, we consider the two-sample comparison problem for panel count data and some nonparametric procedures are discussed. The topic of Section 9.4 is regression analysis of general panel count data using the proportional means model defined in (9.3). Section 9.5 gives some bibliographic notes and discusses some issues and open problems about the analysis of panel count data that are not treated in the previous sections. In this chapter, we assume independent censorship, i.e., (9.4) holds, and the observation process including the number of observations and observation times is independent of the underlying counting process $N^*(t)$ completely or given covariates.

9.2 Nonparametric Estimation of Mean Functions

Consider a follow-up study that involves n independent subjects from a homogeneous population and each subject gives rise to a counting process $N_i(t)$

defined as $N(t)$ in the previous section. Define $\mu(t) = E[N_i(t)]$, the mean function of the processes N_i's. For the ith individual, let $0 < t_{i,1} < \cdots < t_{i,m_i}$ denote the observation time points and $n_{i,j} = N_i(t_{i,j})$, the observed value of $N_i(t)$ at time $t_{i,j}$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. That is, the observed data are

$$\{ \, (\, t_{i,j}, n_{i,j} \,) \; ; \; j = 1, \ldots, m_i, i = 1, \ldots, n \, \} \, .$$

This section discusses two estimators of $\mu(t)$. One is the nonparametric maximum likelihood estimator derived under the non-homogeneous Poisson assumption on the $N_i(t)$'s. The other is the isotonic regression estimator, which is based on an idea similar to that behind weighted least squares regression. The isotonic regression estimator could also be seen as a generalization of the simple sample mean estimator. The two estimators were first studied by Wellner and Zhang (2000) and Sun and Kalbfleisch (1995), respectively.

9.2.1 Nonparametric Maximum Likelihood Estimator

To construct a nonparametric estimator of $\mu(t)$, we first consider the full likelihood approach. Assume that the $N_i(t)$'s are non-homogeneous Poisson processes. Then the log full likelihood function is proportional to

$$l(\mu) \, = \, \sum_{i=1}^{n} \sum_{j=1}^{m_i} (n_{i,j} \, - \, n_{i,j-1}) \, \log[\mu(t_{i,j}) \, - \, \mu(t_{i,j-1})] \, - \, \sum_{i=1}^{n} \mu(t_{i,m_i}) \, ,$$

where $t_{i,0} = 0$ and $n_{i,0} = 0$, and one can estimate $\mu(t)$ by maximizing $l(\mu)$. Let $s_1 < ... < s_m$ denote the ordered distinct observation times in the set $\{ \, t_{i,j} \, ; \; j = 1, ..., m_i, \; i = 1, ..., n \, \}$. Also let $b_l = \sum_{i=1}^{n} I(t_{i,m_i} = s_l)$ for $l = 1, ..., m$ and

$$\tilde{n}_{l,l'} \, = \, \sum_{i=1}^{n} \sum_{j=1}^{m_i} (n_{i,j} \, - \, n_{i,j-1}) \, I(t_{i,j} \, = \, s_l, t_{i,j-1} \, = \, s_{l'}) \, ,$$

for $0 \leq l' < l \leq m$, where $s_0 = 0$. Then the log likelihood function can be rewritten as

$$l(\mu) \, = \, \sum_{l'=0}^{m-1} \, \sum_{l=l'+1}^{m} \, \tilde{n}_{l,l'} \, \log[\mu(s_l) \, - \, \mu(s_{l'})] \, - \, \sum_{l=1}^{m} b_l \, \mu(s_l) \, . \qquad (9.5)$$

It is apparent that only the values of $\mu(t)$ at the s_l's can be estimated and we can define the nonparametric maximum likelihood estimator (NPMLE) of $\mu(t)$, denoted by $\hat{\mu}_F(t)$, as the non-decreasing step function with possible jumps only at the s_l's that maximizes (9.5). Thus the maximization of $l(\mu)$ over functions $\mu(t)$ becomes maximizing $l(\boldsymbol{\mu})$ over m-dimensional parameter vectors $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$ with $\mu_1 \leq ... \leq \mu_m$, where $\mu_l = \mu(s_l)$, $l = 1, ..., m$. Of course, other definitions for $\hat{\mu}_F(t)$ between the s_l's can be used too. Also it can be easily seen that there is no closed solution for the maximizer of $l(\boldsymbol{\mu})$.

For the determination of $\hat{\mu}_F(t)$, for $l = 1, ..., m$, define

$$\phi_l(\boldsymbol{\mu}) = \frac{\partial l(\boldsymbol{\mu})}{\partial \mu_l} \ , \ \phi_{ll}(\boldsymbol{\mu}) = \frac{\partial^2 l(\boldsymbol{\mu})}{\partial \mu_l^2} \ .$$

Also define

$$\Delta_{l,l'}(\boldsymbol{\mu}) = \frac{\sum_{j=l'}^{l} [\phi_j(\boldsymbol{\mu}) - \mu_j \, \phi_{jj}(\boldsymbol{\mu})]}{\sum_{j=l'}^{l} [-\phi_{jj}(\boldsymbol{\mu})]} \ ,$$

$1 \leq l' \leq l \leq m$. Let $\hat{\mu}_{F,l} = \hat{\mu}_F(s_l)$, $l = 1, ..., m$. By using the Fenchel duality theorem, it can be shown that $\hat{\boldsymbol{\mu}}_F = (\hat{\mu}_{F,1}, ..., \hat{\mu}_{F,m})$ satisfies

$$\sum_{l=1}^{m} \phi_l(\hat{\boldsymbol{\mu}}_F) \, \hat{\mu}_{F,l} = 0$$

and

$$\sum_{j=l}^{m} \phi_l(\hat{\boldsymbol{\mu}}_F) \leq 0$$

for all $l = 1, ..., m$ (Wellner and Zhang, 2000). From these, Wellner and Zhang (2000) give the following iterative convex minorant algorithm. Let $\epsilon > 0$ be a prespecified number.

Step 1. Choose an initial estimator $\boldsymbol{\mu}^{(0)} = (\mu_1^{(0)}, ..., \mu_m^{(0)})$.

Step 2. At the kth iteration, obtain the updated estimator by

$$\mu_l^{(k)} = \max_{j' \leq l} \min_{j \geq l} \Delta_{j,j'}(\boldsymbol{\mu}^{(k-1)}) \ , l = 1, ..., m \ ,$$

where $\boldsymbol{\mu}^{(k-1)} = (\mu_1^{(k-1)}, ..., \mu_1^{(k-1)})$ denotes the estimator from the $(k-1)$th iteration.

Step 3. If

$$\left| \sum_{l=1}^{m} \phi_l(\boldsymbol{\mu}^{(k)}) \, \mu_l^{(k)} \right| > \epsilon$$

or

$$\max_{1 \leq l \leq m} \sum_{j=l}^{m} \phi_l(\boldsymbol{\mu}^{(k)}) > \epsilon \ ,$$

return to step 2. Otherwise stop and set $\hat{\mu}_{F,l} = \mu_l^{(k)}$.

For the initial estimator, one can use the sample mean of available observations at each observation time point. Note that although the algorithm described above works well in many applications, sometimes the resulting estimator may not be the globe maximizer. Wellner and Zhang (2000) provide more discussion on this.

9.2.2 Isotonic Regression Estimator

The NPMLE relies on the non-homogeneous Poisson assumption and does not have a closed form. In this subsection, an estimator is discussed that does not need the Poisson assumption and can be easily determined.

To describe the isotonic regression estimator (IRE), we first consider a simple situation where $t_{i,j} = s_j$ for all $i = 1, ..., n$ and $j = 1, ..., m_i$ with $m_i \leq m$. That is, all subjects have the same observation time points except that the numbers of observations may be different. This can be the case in a follow-up study with prespecified observation time points and in which all subjects follow the prespecified observation schedule except that some may drop out of the study early. For this situation, we can use the Nelson-Aalen estimator

$$\int_0^t \frac{\sum_{i=1}^n d\,N_i(s)}{\sum_{i=1}^n I(s \leq t_{i,m_i})}$$

proposed for recurrent event data (Andersen et al., 1993). At time s_l, it gives an estimator

$$\sum_{j=1}^{l} \frac{\sum_{i=1}^n I(s_j \leq t_{i,m_i})[N_i(s_j) - N_i(s_{j-1})]}{\sum_{i=1}^n I(s_j \leq t_{i,m_i})}$$

of $\mu(s_l)$, which is the sample mean of observed values of the $N_i(s_l)$'s from subjects still under study.

For general situations where subjects may not have identical observation times, this Nelson-Aalen estimator is not available. However, we can still define the sample mean at each time point s_l based on available observations. But, unlike the simple situation above, this approach may give an estimator that does not share the non-decreasing property of $\mu(t)$. To fix this, let w_l and $\bar{n}_l$ denote the number and mean value respectively of observations made at s_l, $l = 1, ..., m$. The IRE, denoted by $\hat{\boldsymbol{\mu}}_I = (\hat{\mu}_{I,1}, ..., \hat{\mu}_{I,m})$, of $\boldsymbol{\mu}$ is defined as $\boldsymbol{\mu}$ that minimizes the weighted sum of squares

$$\sum_{l=1}^{m} w_l\,(\,\bar{n}_l - \mu_l\,)^2 \tag{9.6}$$

subject to the order restriction $\mu_1 \leq \cdots \leq \mu_m$ (Sun and Kalbfleisch, 1995). This estimator is the isotonic regression of $\{\bar{n}_1, ..., \bar{n}_m\}$ with weights $\{w_1, ..., w_m\}$ (Robertson et al., 1988). Obviously if $\bar{n}_1 \leq ... \leq \bar{n}_m$, $\hat{\mu}_{I,l} = \bar{n}_l$, $l = 1, ..., m$, and for the simple situation discussed above, the IRE reduces to the Nelson-Aalen estimator given above. Given $\hat{\boldsymbol{\mu}}_I$, the IRE of $\mu(t)$ denoted by $\hat{\mu}_I(t)$ can be defined as the non-decreasing step function with possible jumps only at the s_l's and $\hat{\mu}_I(s_l) = \hat{\mu}_{I,l}$, $l = 1, ..., m$.

The IRE $\hat{\boldsymbol{\mu}}_I$ has a closed form given by

$$\hat{\mu}_{I,l} = \max_{r \leq l} \min_{s \geq l} \frac{\sum_{v=r}^{s} w_v\,\bar{n}_v}{\sum_{v=r}^{s} w_v} = \min_{s \geq l} \max_{r \leq l} \frac{\sum_{v=r}^{s} w_v\,\bar{n}_v}{\sum_{v=r}^{s} w_v}$$

Table 9.1. Observed numbers of loss of feedwater flow from 30 nuclear plants

Observation time t_i (in years) and observed number n_i											
Plant	t_i	n_i	Plant	t_i	n_i	Plant	t_i	n_i	Plant	t_i	n_i
1	15	4	9	4	13	17	2	11	25	1	1
2	12	40	10	3	4	18	2	1	26	3	10
3	8	0	11	4	27	19	2	0	27	2	5
4	8	10	12	4	14	20	1	3	28	4	16
5	6	14	13	4	10	21	1	5	29	3	14
6	5	31	14	2	7	22	1	6	30	11	58
7	5	2	15	3	4	23	5	35			
8	4	4	16	3	3	24	3	12			

using the max-min formula (Barlow et al., 1972, Robertson et al., 1988). In practice, a number of algorithms such as the pool adjacent violators and the up-and-down algorithms can be used to determine $\hat{\boldsymbol{\mu}}_I$.

Suppose that each subject is observed only once as in cross-sectional or some reliability studies. That is, $m_i = 1$, $i = 1, ..., n$. In this case, it can be shown that the two estimators $\hat{\boldsymbol{\mu}}_F$ and $\hat{\boldsymbol{\mu}}_I$ are actually identical (Sun and Kalbfleisch , 1995). Furthermore, if $N_i(t)$ is defined from a survival process, then we have case I interval-censored data and the IRE gives the maximum likelihood estimator of a distribution function discussed in Section 3.2.

9.2.3 Two Examples

This subsection considers two illustrative examples. The first example concerns a simple set of current status panel count data and provides a direct look at how the IRE estimates the mean function. The second example deals with a set of general panel count data.

Table 9.1 presents a set of panel count data arising from a reliability study on the loss of feedwater flow over 30 nuclear plants. The data are reproduced from Gaver and O'Muircheartaigh (1987) and Sun and Kalbfleisch (1995) and consist of the observation time (one per plant) and the corresponding observed number of losses of feedwater flow for each nuclear plant. There are a total of 10 different observation time points ($m = 10$). Assume that the numbers of the losses of feedwater flow for all 30 nuclear plants follow the same counting process. To determine the IRE of the mean or average number of losses of feedwater flow based on the observed data, we first calculate the sample mean of the numbers of the observed losses of feedwater flow ($\bar{n}_l$) at each observation time point. Figure 9.1 presents the IRE of the average number of losses of feedwater flow given by the max-min formula. Note that for the current situation, the NPMLE and IRE are identical. For comparison and understanding the IRE, Figure 9.1 also includes the sample means of the numbers of observed losses ($\bar{n}_l$ vs s_l). It can be clearly seen that the IRE is obtained by pooling the $\bar{n}_l$'s according to the order restriction.

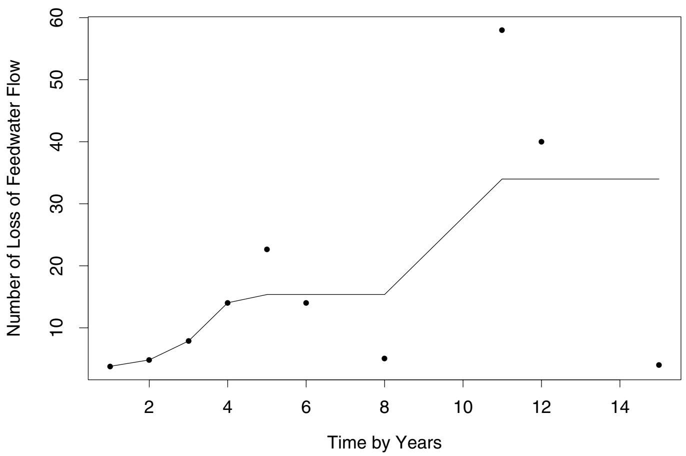

Fig. 9.1. IRE of the average number of losses of feedwater flow.

As a second example, consider data set IV of Appendix A, a set of panel count data arising from the National Cooperative Gallstone Study. This is a 10-year, multicenter, double-blinded, placebo-controlled clinical trial on the use of the natural bile acid chenodeoxycholic acid, cheno, for the dissolution of cholesterol gallstones. In the original study, a total of 916 patients were randomized into each of three treatments, placebo, low dose, and high dose, and were treated for up to two years. One of the primary objectives of the study was to assess the impact of the treatments on the incidence of digestive symptoms commonly associated with gallstone disease. The symptoms range from milder episodes of nausea/vomiting, dyspepsia, and diarrhea to more severe episodes of digestive colic, i.e., severe pain, and cholecystitis, i.e., digestive obstruction. The data set is reproduced from Thall and Lachin (1988) and contains the observed information on the incidence of nausea from the first year follow-up on 111 patients with floating gallstones in high-dose (63) and placebo (48) groups.

Nausea is an unpleasant sensation vaguely referred to the epigastrium and abdomen, often culminating in vomiting. It is very commonly associated with gallstone disease and it is important to the investigators to determine whether there exists a significant difference between the incidence of nausea for the patients in the two groups. It was hypothesized that any treatment effect should be observed shortly after patients achieved maximal dose (usually by three months). The effect might later begin to dissipate. Thus only the first year data are studied.

During the study, the patients were scheduled to return for clinic observations at 1, 2, 3, 6, 9, and 12 months during the first year follow-up. At each

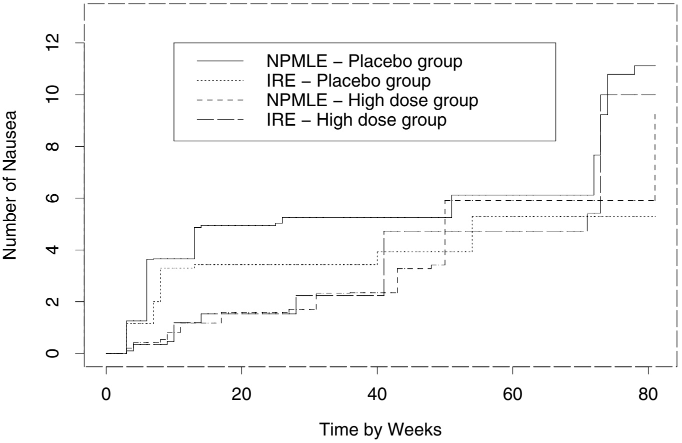

Fig. 9.2. Estimates of the average cumulative counts of episodes of nausea.

visit, they were asked to report the total number of each type of symptom that had occurred between successive visits such as the number of the incidences of nausea. That is, the observed data include actual visit times and the numbers of the incidences or occurrences of nausea between the visits. As expected, actual visit or observation times differ from patient to patient. For example, first observation times ranged from 3 to 9 weeks, and some patients dropped out of the study early.

To estimate the average cumulative numbers of the occurrences of nausea for the patients in the placebo and high-dose groups, the NPMLE and IRE are obtained and displayed in Figure 9.2. They suggest that the patients in the placebo group seem to have higher incidence of nausea than those in the high-dose group over the first 40 weeks. Most of this difference seems due to an early difference over the first 10 weeks. After 40 weeks, the incidence of nausea for the patients in the high-dose group seems to catch up that for those in the placebo group. A possible reason for this is that the treatment, cheno, may only have short-term effects.

It is interesting to note that for the patients in the high-dose group, the NPMLE and IRE are quite close to each other, especially for the period of the first 40 weeks. In contrast, the two estimates for those in the placebo group differ and the NPMLE gives a higher estimate of the incidence of nausea. To explain this difference, we calculate the empirical estimates of the rates of the incidence of nausea for the two groups given by

$$d\hat{\mu}_e(t) = \frac{1}{\sum_{i=1}^n I(t \le t_{i,m_i})} \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{n_{i,j} - n_{i,j-1}}{t_{i,j} - t_{i,j-1}} I(t_{i,j-1} < t \le t_{i,j})$$

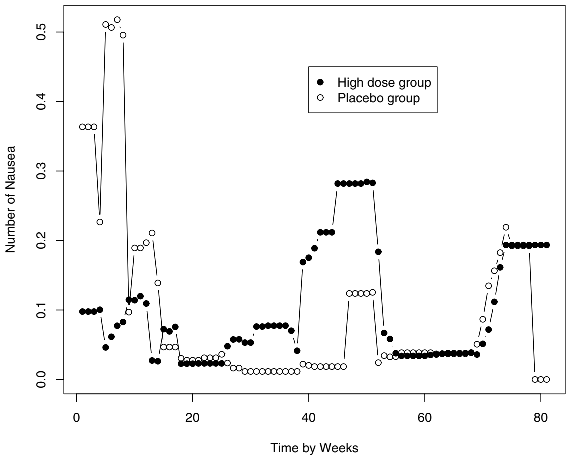

Fig. 9.3. Empirical estimates of the rates of the incidence of episodes of nausea.

(Thall and Lachin, 1988), the average of estimated individual rate functions. The two estimates are presented in Figure 9.3. The figure shows that for the patients in the placebo group, the incidence of nausea is relatively quite higher over the initial period, then becomes lower subsequently. This suggests that the NPMLE is more influenced by the earlier and higher episodes of nausea than the IRE, which seems to be consistent with the nature of the IRE.

9.2.4 Discussion

The IRE $\hat{\boldsymbol{\mu}}_I$ also can be derived under the non-homogeneous Poisson process assumption. For this, note that ignoring the dependency of $\{N_i(t_{ij}), j = 1, ..., m_i\}$ for each i, one can construct a pseudo log likelihood function

$$l_p(\mu) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} [\, n_{i,j} \, \log \mu(t_{i,j}) \, - \, \mu(t_{i,j}) \,] = \sum_{l=1}^{m} w_l \, (\, \bar{n}_l \, \log \mu_l \, - \, \mu_l \,) \quad (9.7)$$

for μ under the non-homogeneous Poisson assumption. It can be shown that the maximization of $l_p(\mu)$ is equivalent to the minimization of (9.6) (Robertson et al. 1988; Wellner and Zhang, 2000).

In comparing the two estimators of $\mu(t)$ discussed above, it is easy to see from (9.5) and (9.7) that the NPMLE could be more efficient than the IRE. Wellner and Zhang (2000) studied this by simulation and suggested that this is true for both non-homogeneous Poisson processes and some other counting processes. A disadvantage of the NPMLE is that its implementation is much more involved in terms of programming and requires much more computing time than that of the IRE. In general, the latter provides a general idea about the shape of the mean function $\mu(t)$, especially for the case where the number of observations for each subject is small. The former should be used, for example, if the non-homogeneous Poisson assumption seems reasonable.

Wellner and Zhang (2000) investigated the asymptotic properties of the NPMLE and IRE. In particular, they prove that under some regularity conditions, both estimators are consistent in L_2 and for fixed t, $n^{1/3}\,[\hat{\mu}_F(t)\,-\,\mu(t)]$ and $n^{1/3}\,[\hat{\mu}_I(t)\,-\,\mu(t)]$ converge in distribution to the maximum point of a two-sided Brownian motion process multiplied by some constants. Discussion about this limit distribution can be found in Groeneboom and Wellner (2001). We remark that these asymptotic results do not rely on the non-homogeneous Poisson assumption although the NPMLE is derived under this assumption.

As discussed above, simply from the non-homogeneous Poisson process point of view, the NPMLE treats the dependence of $\{\,n_{i,j}\,,\,j\,=\,1,...,m_i\}$ as it is, while the IRE completely ignores this dependence. One could model or make some assumptions about this dependence to provide an intermediate approach. Zhang and Jamshidian (2003) proposed such an approach by assuming that given a latent variable b_i, $E[N_i(t)|b_i]\,=\,b_i\,\mu(t)$ and $\{\,n_{i,j}\,,\,j=1,...,m_i\}$ are independent, $i\,=\,1,...,n$. Furthermore, by assuming that the b_i's follow a gamma distribution, they developed an EM algorithm for estimation of $\mu(t)$. The efficiency of the resulting estimator may be between the NPMLE and IRE, but as the NPMLE, its determination is still much more involved than that of the IRE although is simpler than that of the NPMLE. In addition, the theoretical investigation of its properties is much harder than that of the NPMLE and IRE.

Instead of direct estimation of the mean function $\mu(t)$, which has to take into account the monotonic property of $\mu(t)$, an alternative is to estimate the rate function $d\,\mu(t)$ first and then to estimate $\mu(t)$ by the integral of the rate estimator. Thall and Lachin (1988) considered this approach by estimating $d\,\mu(t)$ using the empirical estimate $d\,\mu_e(t)$ given in Section 9.2.3.

9.3 Nonparametric Comparison of Mean Functions

This section considers the same setup as in the previous section but supposes that study subjects come from two different treatment groups with z_i being the group indicator, $i\,=\,1,...,n$. Define $\mu_i(t)\,=\,\mu(t\,;\,z_i)\,=\,E[N_i(t)|z_i]$, $i\,=\,1,...,n$. The goal is to test the hypothesis $H_0\,:\,\mu_1(t)\,=\,...\,=\,\mu_n(t)$. That is, the two treatment groups have the same mean functions.

9.3.1 A Generalized Score Test Procedure

To motivate the test statistic for H_0, first consider a simple situation where each subject is observed only once and the mean function $\mu_i(t)$ satisfies the model (9.3) with β representing the group difference. In this case, the hypothesis H_0 is equivalent to $\beta = 0$ and the log likelihood function of μ_0 and β is proportional to

$$l(\mu_0, \beta) = \sum_{i=1}^{n} [\, n_{i,1} \log \mu_0(t_{i,1}) + n_{i,1}\, z_i\, \beta - \mu_0(t_{i,1}) \exp(z_i\, \beta) \,]$$

using the notation defined in the previous section. A natural statistic for testing $\beta = 0$ is the score statistic from the log likelihood function $l(\mu_0, \beta)$, which has the form

$$\left. \frac{\partial l(\mu_0, \beta)}{\partial \beta} \right|_{\beta=0} = \sum_{i=1}^{n} z_i \, [\, n_{i,1} - \mu_0(t_{i,1}) \,]$$

with $\mu_0(t)$ replaced by its estimator.

For general panel count data, let $\hat{\mu}_I(t)$ denote the IRE of $\mu_i(t) = \mu_0(t)$ under H_0 given in Section 9.2. By generalizing the score statistic given above, for testing H_0, one can use the test statistic

$$U = \sum_{i=1}^{n} z_i \sum_{j=1}^{m_i} [\, n_{i,j} - \hat{\mu}_I(t_{i,j}) \,] \, . \tag{9.8}$$

Suppose that $z_i = 0$ or 1 and let $\hat{\mu}_I^{(u)}(t)$, $\{\, s_l^{(u)} \,\}$ and $\{\, w_l^{(u)} \,\}$ be defined as $\hat{\mu}_I(t)$, $\{\, s_l \,\}$ and $\{\, w_l \,\}$ in the previous section, but based only on subjects with $z_i = u$, $u = 0, 1$. Then the test statistic U can be rewritten as

$$U = \int w^{(1)}(t) \, [\, \hat{\mu}_I^{(1)}(t) - \hat{\mu}_I(t) \,] \, d\bar{N}^{(1)}(t) \, ,$$

where $w^{(1)}(t)$ is a step function with jumps only at the $s_l^{(1)}$'s and $w^{(1)}(s_l^{(1)}) = w_l^{(1)}$ and $\bar{N}^{(1)}(t) = \sum_l I(t \geq s_l^{(1)})$. That is, U is the integrated weighted difference between an individual group estimator $\hat{\mu}_I^{(1)}(t)$ and the overall estimator $\hat{\mu}_I(t)$.

Sun and Kalbfleisch (1993) and Sun and Fang (2003) studied the statistic U for current status panel count data and general panel count data, respectively. In particular, they show that under some regularity conditions and H_0, as $n \to \infty$, $n^{-1/2}\, U$ asymptotically has a normal distribution with mean zero and variance that can be consistently estimated by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left[(z_i - \bar{z}) \sum_{j=1}^{m_i} (n_{i,j} - \hat{\mu}_I(t_{i,j})) \right]^2 \, ,$$

where $\bar{z} = \sum_{i=1}^{n} z_i / n$. Hence one can test the hypothesis H_0 using the statistic $U^* = U \,/ (n^{1/2}\, \hat{\sigma})$ based on the standard normal distribution.

One condition required for the asymptotic distribution given above is that the z_i's can be treated as independent and identically distributed random variables with finite variance when n is large. This may seem restrictive. But it is often reasonable for large randomized clinical trials. Also Sun and Kalbfleisch (1993) and Sun and Fang (2003) show by simulation that the result holds as long as the z_i's are a fixed random permutation of a sequence of constants such as a sequence of zero and one. The result above also requires that the z_i's are independent of the observations, which usually holds in most studies. It should be noted that some randomness on the z_i's is necessary for the validity of the statistic U. To see this, consider an extreme situation where the mean function satisfies model (9.3) and the z_i's are a sequence of zero followed by a sequence of one with $t_{i_1,m_{i_1}} \leq t_{i_2,m_{i_2}}$ for $i_1 < i_2$. In this case, the information about the baseline mean function $\mu_0(t)$ and β cannot be separated.

To illustrate the statistical method presented here, consider the set of panel count data discussed in Section 9.2.3 and presented in data set IV of Appendix A. Define $z_i = 1$ for patients in the placebo group and $z_i = 0$ otherwise. Let $N_i(t)$ denote the cumulative number of occurrences of nausea up to weeks t for the ith patient, $i = 1, ..., 111$. The application of the test procedure gives $U_n^* = 0.7328$ based on the data observed up to 58 weeks. This gives a p-value of 0.4637 according to the standard normal distribution. The result suggests that the average incidences of nausea did not differ significantly overall between the patients in the placebo and high-dose groups. In summary, although Figures 9.2 and 9.3 indicate that there exists some difference between the groups over different time periods, there seems no overall difference between the incidences of nausea of the two groups.

9.3.2 Discussion

An obvious alternative to the statistic U is to replace the IRE by the NPMLE in (9.8). Another alternative is to use the statistic

$$\int w^{(1)}(t)\,[\,\hat{\mu}_I^{(1)}(t) - \hat{\mu}_I(t)\,]^2\, d\,\bar{N}^{(1)}(t)$$

or

$$\int W(t)\,|\,\hat{\mu}_I^{(1)}(t) - \hat{\mu}_I^{(0)}(t)\,|\, d\,\bar{N}(t)$$

for testing H_0, where $W(t)$ is a weight process and $\bar{N}(t) = \sum_j I(t \geq s_j)$. These alternative test statistics could be more efficient than U. However, their asymptotic null distributions are unknown.

To test H_0, in addition to the techniques given above that directly compare the underlying mean functions, one could employ indirect approaches. For example, Thall and Lachin (1988) compared the observed numbers of the event over different time intervals. Specifically, they first partitioned the entire study period into K fixed, consecutive intervals and transformed the observed information on each subject into a K-dimensional vector. The two-sample comparison was then performed by using the two-sample test given in Wei and Lachin (1984) for K-variate nonnegative-valued random vectors. It is apparent that the test result could depend on the selection of the number of intervals and the intervals themselves.

Sometimes one may need to compare more than two mean functions. Suppose that study subjects come from p different populations and let $\mu^{(u)}(t)$ denote the mean function of counting processes arising from the subject in the uth population, $u = 1, ..., p$. To test the hypothesis $H_0' : \mu^{(1)}(t) = ... = \mu^{(p)}(t)$, let $\hat{\mu}^{(u)}(t)$ denote the NPMLE or IRE of $\mu^{(u)}(t)$ based only on the subjects within the uth population and $\hat{\mu}(t)$ the NPMLE or IRE of the common mean function under H_0'. Also let $w^{(u)}(t)$ and $\bar{N}^{(u)}(t)$ be defined as $w^{(1)}(t)$ and $\bar{N}^{(1)}(t)$ before, but based on the subjects from the uth population. Then for the test of H_0', one can use the statistic $\boldsymbol{U} = (U_1, ..., U_p)'$, where

$$U_u = \int w^{(u)}(t) \, [\, \hat{\mu}^{(u)}(t) - \hat{\mu}(t) \,] \, d \bar{N}^{(u)}(t) \, ,$$

$u = 1, ..., p$. Alternatively, one can apply the statistic $\boldsymbol{U}^* = (U_2^*, ..., U_p^*)'$, where

$$U_u^* = \int w^{(u)}(t) \, [\, \hat{\mu}^{(u)}(t) - \hat{\mu}^{(1)}(t) \,] \, d \bar{N}^{(u)}(t) \, ,$$

$u = 2, ..., p$, assuming that the first population represents a control group. For the use of either $\boldsymbol{U}$ or $\boldsymbol{U}^*$, however, one needs to derive its null asymptotic distribution.

9.4 Regression Analysis of Panel Count Data

This section deals with regression analysis of panel count data. Let the $N_i(t)$, $\mu(t\,;\, \boldsymbol{Z}_i)$, $t_{i,j}$'s, $n_{i,j}$'s, and s_l's be defined as in the previous sections. Suppose that for subject i, there exists a p-dimensional vector of covariates denoted by $\boldsymbol{Z}_i$, assumed to be time-independent, and that one observes data

$$\{ \, (\, t_{i,j}, n_{i,j}, \boldsymbol{Z}_i \,) \; ; \; j = 1, \ldots, m_i, i = 1, \ldots, n \, \} \, .$$

We assume that the mean function $\mu(t\,;\, \boldsymbol{Z}_i)$ satisfies model (9.3) and the main goal is to make inferences about the regression parameters $\boldsymbol{\beta}$.

Following the discussion in Section 9.2, a natural inference approach is to treat the $N_i(t)$'s as non-homogeneous Poisson processes. In this case, using the notation defined before, the log likelihood functions given in (9.5) and (9.7) are

$$l(\mu_0, \boldsymbol{\beta}) = \sum_{l'=0}^{m-1} \sum_{l=l'+1}^{m} \tilde{n}_{l,l'} \, \log[\mu_0(s_l) - \mu_0(s_{l'})] - \sum_{l=1}^{m} b_l(\boldsymbol{\beta}) \, \mu_0(s_l)$$

$$+ \sum_{i=1}^{n} n_{i,m_i} \, \boldsymbol{Z}_i' \boldsymbol{\beta}$$

and

$$l_p(\mu_0, \boldsymbol{\beta}) \,=\, \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[n_{i,j} \, \log \mu_0(t_{i,j}) \,+\, n_{i,j} \, \boldsymbol{Z}_i' \boldsymbol{\beta} \,-\, \mu_0(t_{i,j}) \, \exp(\boldsymbol{Z}_i' \boldsymbol{\beta}) \right] \,,$$

respectively, where $b_l(\boldsymbol{\beta}) \,=\, \sum_{i=1}^{n} I(t_{i,m_i} \,=\, s_l) \, \exp(\boldsymbol{Z}_i' \boldsymbol{\beta})$. In the following, we discuss two approaches. One is based on the pseudo log likelihood function $l_p(\mu_0, \boldsymbol{\beta})$ (Zhang, 2002; Wellner et al., 2004) and the other on estimating equations (Sun and Wei, 2000). Although the log likelihood function $l(\mu_0, \boldsymbol{\beta})$ may seem more attractive than $l_p(\mu_0, \boldsymbol{\beta})$, the unknown properties of estimates resulting from it and its complexity limit its applicability.

9.4.1 A Non-homogeneous Poisson Process Approach

This subsection discusses the non-homogeneous Poisson process approach to estimation of parameters. Let the w_l's and $\bar{n}_l$'s be defined as before. Define

$$\bar{a}_l(\boldsymbol{\beta}) \,=\, \frac{1}{w_l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \exp(\boldsymbol{Z}_i' \boldsymbol{\beta}) \, I(t_{i,j} = s_l)$$

and

$$\bar{b}_l(\boldsymbol{\beta}) \,=\, \frac{1}{w_l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} n_{i,j} \, \boldsymbol{Z}_i' \boldsymbol{\beta} \, I(t_{i,j} = s_l)$$

for given $\boldsymbol{\beta}$, $l \,=\, 1, ..., m$. Then the pseudo log likelihood function $l_p(\mu_0, \boldsymbol{\beta})$ can be rewritten as

$$l_p(\mu_0, \boldsymbol{\beta}) \,=\, \sum_{l=1}^{m} w_l \left[\bar{n}_l \, \log \mu_0(s_l) \,-\, \bar{a}_l(\boldsymbol{\beta}) \, \mu_0(s_l) \,+\, \bar{b}_l(\boldsymbol{\beta}) \right]$$

and one can estimate $\boldsymbol{\beta}$ as well as $\mu_0(t)$ by maximizing $l_p(\mu_0, \boldsymbol{\beta})$.

As in Section 9.2, only the values of $\mu_0(t)$ at the s_l's can be estimated. Let $\hat{\mu}_0(t)$ and $\hat{\boldsymbol{\beta}}$ denote the estimators of $\mu_0(t)$ and $\boldsymbol{\beta}$ defined above with $\hat{\mu}_0(t)$ being a non-decreasing step function with possible jumps only at the s_l's. The determination of $\hat{\mu}_0(t)$ and $\hat{\boldsymbol{\beta}}$ is equivalent to maximizing $l_p(\mu_0, \boldsymbol{\beta}) \,=\, l_p(\boldsymbol{\mu}, \boldsymbol{\beta})$ over the $(m \,+\, p)$ unknown parameters $\boldsymbol{\mu} \,=\, (\mu_1, ..., \mu_m)$ and $\boldsymbol{\beta}$ with $\mu_1 \leq ... \leq \mu_m$, where $\mu_l \,=\, \mu_0(s_l)$, $l \,=\, 1, ..., m$.

To maximize $l_p(\boldsymbol{\mu}, \boldsymbol{\beta})$, one can use a two-step iterative algorithm that maximizes l_p over $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ alternatively. For fixed $\boldsymbol{\beta}$, note that the maximization of l_p over $\boldsymbol{\mu}$ is equivalent to maximizing

$$\sum_{l=1}^{m} w_l \, \bar{a}_l(\boldsymbol{\beta}) \left(\frac{\bar{n}_l}{\bar{a}_l(\boldsymbol{\beta})} \, \log \mu_l \,-\, \mu_l \right) \,,$$

which is similar to the log likelihood function given in (9.7). This shows that with $\boldsymbol{\beta}$ given, the $\hat{\mu}_0(s_l)$'s are the isotonic regression estimator of $\{\bar{n}_1/\bar{a}_1(\boldsymbol{\beta}), ..., \bar{n}_m/\bar{a}_m(\boldsymbol{\beta})\}$ with weights $\{w_1\bar{a}_1(\boldsymbol{\beta}), ..., w_m\bar{a}_m(\boldsymbol{\beta})\}$. Thus they have the closed form

$$\hat{\mu}_{\boldsymbol{\beta}}(s_l) = \max_{r \le l} \min_{s \ge l} \frac{\sum_{v=r}^{s} w_v\, \bar{n}_v}{\sum_{v=r}^{s} w_v \bar{a}_v(\boldsymbol{\beta})} = \min_{s \ge l} \max_{r \le l} \frac{\sum_{v=r}^{s} w_v\, \bar{n}_v}{\sum_{v=r}^{s} w_v \bar{a}_v(\boldsymbol{\beta})}$$

given by the max-min formula of the isotonic regression estimate (Barlow et al., 1972, Robertson et al., 1988).

For given $\mu_0(t)$ or $\boldsymbol{\mu}$, one can simply use the Newton-Raphson algorithm for estimation of $\boldsymbol{\beta}$. It can be easily shown that the pseudo log likelihood function l_p is a concave function of $\boldsymbol{\beta}$ for given $\mu_0(t)$ and its value increases after each iteration (Zhang, 2002). For the convergence criterion for the two-step algorithm given above, one can compare the relative absolute change of either the log likelihood function l_p between two successive estimators of $\mu_0(t)$ and $\boldsymbol{\beta}$ or the difference between the two successive estimators.

In general, the two-step iterative algorithm described above is robust and seems always to converge (Zhang, 2002). Under some regularity conditions, Zhang (2002) shows that the estimators of $\mu_0(t)$ and $\boldsymbol{\beta}$ given by l_p are consistent in L_2 even if the $N_i(t)$'s are not Poisson processes.

9.4.2 An Estimating Equation Approach

The inference in Section 9.4.1 uses a conditional approach in the sense that it conditions on observation times or treats them as fixed. Sometimes it may be better to directly model them together with the counting process of interest and to make unconditional inferences about covariate effects. This subsection discusses methods of this type that make use of estimating equations and model the observation process marginally as well as the process of interest. The method also deals with the censoring or follow-up time, which may be related to both the counting process of interest and the observation process and is not considered in the non-homogeneous Poisson process approach. Also note that formal inference about $\boldsymbol{\beta}$ cannot be carried out using the non-homogeneous Poisson process approach because no asymptotic distribution or variance for $\hat{\boldsymbol{\beta}}$ is available although one could apply the bootstrap procedure. In contrast, formal inference procedures are given below.

9.4.2.1 Models

For subject i, suppose that there exists a censoring or follow-up time denoted by C_i and let $\tilde{N}^*(t)$ be the underlying observation process denoting the potential number of observations up to time t on the subject, $i = 1, ..., n$. Then $\tilde{N}_i = \tilde{N}_i^*\{\min(t, C_i)\} = \sum_{j=1}^{m_i} I(t_{i,j} \le t)$ represents the real observation process on the ith subject and $N_i(t)$ is observed only at the time points where $\tilde{N}_i(t)$ jumps, $i = 1, ..., n$. In the following, it is assumed that $[\{N_i^*(t), \tilde{N}_i^*(t), C_i, \boldsymbol{Z}_i\}' \,;\, t \geq 0\,,\, i = 1, ..., n]$ are independent and identically distributed. Also assume that N_i^*, $\tilde{N}_i^*$, C_i and $\boldsymbol{Z}_i$ may be dependent, but given $\boldsymbol{Z}_i$, the quantities N_i^*, $\tilde{N}_i^*$ and C_i are independent.

To model the dependence of $\tilde{N}_i^*$ on $\boldsymbol{Z}_i$, as for N_i^*, it is assumed that the mean function of $\tilde{N}_i^*(t)$ given $\boldsymbol{Z}_i$ has the form

$$\tilde{\mu}_i(t\,;\,\boldsymbol{Z}_i) \,=\, E\,[\,\tilde{N}_i^*(t)\,|\,\boldsymbol{Z}_i\,] \,=\, \tilde{\mu}_0(t)\,\exp(\boldsymbol{Z}_i'\,\boldsymbol{\gamma})\,. \tag{9.9}$$

In this model, $\tilde{\mu}_0(t)$, as $\mu_0(t)$, is a completely unspecified function and $\boldsymbol{\gamma}$ is a p-dimensional vector of regression parameters representing the effect of covariates on $\tilde{N}_i^*$. For the follow-up time, suppose that given $\boldsymbol{Z}_i$, the hazard function $\lambda_i^*(t)$ of C_i is given by the PH model

$$\lambda_i^*(t\,;\,\boldsymbol{Z}_i) \,=\, \lambda_0^*(t)\,\exp(\boldsymbol{Z}_i'\,\boldsymbol{\tau})\;, \tag{9.10}$$

where $\lambda_0^*(t)$ is a completely unspecified function and $\boldsymbol{\tau}$ is a p-dimensional vector of regression parameters denoting the effect of covariates on C_i. Note that here C_i is always observable unlike the case of right-censored failure time data. In the following, for simplicity of presentation, it is assumed that the $\boldsymbol{Z}_i$'s are centered around zero. Otherwise, one can simply replace $\boldsymbol{Z}_i$ by $\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_n$, where $\bar{\boldsymbol{Z}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i$.

The following first considers the general situation where all regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ are unknown and need to be estimated. Then the special case where $\boldsymbol{\tau} = 0$ is discussed, which implies that the C_i's are independent of all concerned processes.

9.4.2.2 Estimation of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\tau}$

To describe the estimating equation for regression parameters, first consider a simple situation where $m_i = 1$ and $\boldsymbol{\gamma} = \boldsymbol{\tau} = 0$, $i = 1, ..., n$. That is, one only has current status panel count data and the $\tilde{N}_i^*(t)$'s and C_i's have the same mean and hazard functions, respectively. In this case, under model (9.3), the mean of $\exp(-\,\boldsymbol{Z}_i'\,\boldsymbol{\beta})\,N_i(t_{i,1}) = \exp(-\,\boldsymbol{Z}_i'\,\boldsymbol{\beta})\int N_i(t)\,d\tilde{N}_i(t)$ is independent of i. To estimate β, consider testing model (9.3), for which a natural statistic due to Wilcoxon is

$$U_0^*(\boldsymbol{\beta}) \,=\, \sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{Z}_i \,-\, \boldsymbol{Z}_j)\,\left[\exp(-\boldsymbol{Z}_i'\,\boldsymbol{\beta})\int N_i(t)\,d\tilde{N}_i(t)\right.$$

$$\left. -\,\exp(-\boldsymbol{Z}_j'\,\boldsymbol{\beta})\int N_j(t)\,d\tilde{N}_j(t)\right]$$

$$= 2\,n\sum_{i=1}^n \left[\boldsymbol{Z}_i\,\exp(-\boldsymbol{Z}_i'\,\boldsymbol{\beta})\int N_i(t)\,d\tilde{N}_i(t)\right]$$

given $\boldsymbol{\beta}$. This suggests that one can use the estimating equation $U_0(\boldsymbol{\beta}) = (2\,n)^{-1}\,U_0^*(\boldsymbol{\beta}) = 0$ for estimation of $\boldsymbol{\beta}$.

For cases with $m_i \geq 1$, $i = 1, ..., n$, one can still use this estimating equation and in this case, one has $\int N_i(t)\,d\tilde{N}_i(t) = \sum_{j=1}^{m_i} N_i(t_{i,j})$. It is easy to see that $U_0(\boldsymbol{\beta})$ is an unbiased estimating function under model (9.3) with $\boldsymbol{\gamma} = \boldsymbol{\tau} = 0$.

Now consider the general case where $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ may not be zero. Let $S_0(t) = \exp[-\int_0^t \lambda_0^*(s)\,ds]$ and define

$$d\tilde{M}_i(t) = d\tilde{N}_i(t) - I(C_i \geq t)\,\exp(\boldsymbol{Z}_i'\,\boldsymbol{\gamma})\,d\tilde{\mu}_0(t)\ ,$$

which has mean zero, $i = 1, ..., n$. Then one has

$$\int N_i(t) d\tilde{N}_i(t) = \int N_i(t) d\tilde{M}_i(t) + \int N_i(t)\exp(\boldsymbol{Z}_i'\,\boldsymbol{\gamma})\,I(C_i \geq t) d\tilde{\mu}_0(t)\ ,$$

and under model (9.9), conditional on $\boldsymbol{Z}_i$,

$$E\left[\int N_i(t) d\tilde{N}_i(t)\right] = \exp[\,\boldsymbol{Z}_i'\,(\boldsymbol{\beta}+\boldsymbol{\gamma})]\int \mu_0(t)\,S_i(t)\,d\tilde{\mu}_0(t)\ , \qquad (9.11)$$

where $S_i(t) = P(C_i \geq t) = [S_0(t^-)]^{\exp(\boldsymbol{Z}_i'\,\boldsymbol{\tau})}$ under model (9.10). Equation (9.11) shows that $U_0(\boldsymbol{\beta})$ is biased in the situations considered and needs to be adjusted.

To have an unbiased estimating function similar to $U_0(\boldsymbol{\beta})$, it follows from (9.11) that one should consider the quantity

$$\int N_i(t)\,[S_0(t^-)]^{-\exp(\boldsymbol{Z}_i'\,\boldsymbol{\tau})}\,d\tilde{N}_i(t)$$

instead of $\int N_i(t) d\tilde{N}_i(t)$. Under model (9.10), this quantity has expectation

$$\exp[\,\boldsymbol{Z}_i'\,(\boldsymbol{\beta}+\boldsymbol{\gamma})\,]\int \mu_0(t)\ d\tilde{\mu}_0(t)\ .$$

This motivates the estimating function

$$U(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\tau}) = \sum_{i=1}^n \boldsymbol{Z}_i e^{-\boldsymbol{Z}_i'\,(\boldsymbol{\beta}+\boldsymbol{\gamma})}\int N_i(t)\,[\hat{S}_0(t^-;\boldsymbol{\tau})]^{-\exp(\boldsymbol{Z}_i'\,\boldsymbol{\tau})}\,d\tilde{N}_i(t)$$

for $\boldsymbol{\beta}$ with fixed $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$, where

$$\hat{S}_0(t;\boldsymbol{\tau}) = \exp\left[-\int_0^t \frac{d\,\bar{N}(s)}{\sum_{i=1}^n I(C_i \geq s)\,e^{\boldsymbol{Z}_i'\,\boldsymbol{\tau}}}\right]\ ,$$

$\bar{N}(s) = \sum_{i=1}^n \bar{N}_i(s)$ and $\bar{N}_i(s) = I(C_i \leq s)$. It can be easily shown that asymptotically, $U(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\tau})$ has expectation zero under the true values of the parameters (Sun and Wei, 2000).

To estimate $\boldsymbol{\gamma}$ under model (9.9), a common approach is to use the estimating equation $U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \partial L(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = 0$ (Lawless and Nadeau, 1995), where

$$L(\boldsymbol{\gamma}) = \int \sum_{i=1}^{n} \left[\boldsymbol{Z}_i'\boldsymbol{\gamma} - \log[\sum_{l=1}^{n} I(C_l \geq t)\exp(\boldsymbol{Z}_l'\boldsymbol{\gamma})] \right] d\tilde{N}_i(t) \, .$$

For estimation of $\boldsymbol{\tau}$, one can use the partial likelihood score function

$$U_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \sum_{i=1}^{n} \int \left[\boldsymbol{Z}_i - \frac{\sum_{l=1}^{n} I(C_i \geq t)\, e^{\boldsymbol{Z}_l'\boldsymbol{\tau}}\, \boldsymbol{Z}_l}{\sum_{l=1}^{n} I(C_i \geq t)\, e^{\boldsymbol{Z}_l'\boldsymbol{\tau}}} \right] d\, I(C_i \leq t)$$

(Kalbfleisch and Prentice, 2002). Let $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\tau}}$ denote the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ given by the solutions to $U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = 0$ and $U_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = 0$, respectively. Then one can estimate $\boldsymbol{\beta}$ by the solution, denoted by $\tilde{\boldsymbol{\beta}}$, to $U(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\tau}}) = 0$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\tau}')'$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\boldsymbol{\gamma}}', \tilde{\boldsymbol{\tau}}')'$. Sun and Wei (2000) show that the estimators $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\tau}}$ are consistent and unique. For their asymptotic distributions, let

$$A(\boldsymbol{\theta}) = -\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \, , \; B(\boldsymbol{\gamma}) = -\frac{\partial U_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \, , \; G(\boldsymbol{\tau}) = -\frac{\partial U_{\boldsymbol{\tau}}(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \, , \; P(\boldsymbol{\theta}) = -\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\tau}} \, .$$

Define

$$R(t, \boldsymbol{\theta}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{Z}_i\, e^{-\boldsymbol{Z}_i'(\boldsymbol{\beta}+\boldsymbol{\gamma}-\boldsymbol{\tau})} \int_t^{\infty} \frac{N_i(s)}{[\hat{S}_0(s, \tau)]^{\exp(\boldsymbol{Z}_i'\boldsymbol{\tau})}}\, d\,\tilde{N}_i(s)$$

and

$$S^{(j)}(t, \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} I(C_i \geq t)\, e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}\, \boldsymbol{Z}_i^{(j)} \, ,$$

where $j = 0$ and 1, $\boldsymbol{Z}_i^{(0)} = 1$, and $\boldsymbol{Z}_i^{(1)} = \boldsymbol{Z}_i$, $i = 1, ..., n$. Also define

$$\tilde{a}_i(\boldsymbol{\theta}) = \boldsymbol{Z}_i\, e^{-\boldsymbol{Z}_i'(\boldsymbol{\beta}+\boldsymbol{\gamma})} \int \frac{N_i(t)}{[\hat{S}_0(t, \boldsymbol{\tau})]^{\exp(\boldsymbol{Z}_i'\boldsymbol{\tau})}}\, d\,\tilde{N}_i(t) \, ,$$

$$\tilde{b}_i(\boldsymbol{\theta}) = \int \frac{R(t, \boldsymbol{\theta})}{S^{(0)}(t, \boldsymbol{\tau})} \left[d\,\bar{N}_i(t) - \frac{I(C_i \geq t)\, e^{\boldsymbol{Z}_i'\boldsymbol{\tau}}}{n\, S^{(0)}(t, \boldsymbol{\tau})}\, d\,\bar{N}(t) \right] \, ,$$

$$\tilde{d}_i(\boldsymbol{\gamma}) = \int_0^{\infty} \left[\boldsymbol{Z}_i - \frac{S^{(1)}(t, \boldsymbol{\gamma})}{S^{(0)}(t, \boldsymbol{\gamma})} \right] \left[d\tilde{N}_i(t) - \frac{I(C_i \geq t)\, e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}}{n\, S^{(0)}(t, \boldsymbol{\gamma})}\, d\tilde{N}(t) \right]$$

and

$$\tilde{d}_i(\boldsymbol{\tau}) = \int_0^{\infty} \left[\boldsymbol{Z}_i - \frac{S^{(1)}(t, \boldsymbol{\tau})}{S^{(0)}(t, \boldsymbol{\tau})} \right] \left[d\bar{N}_i(t) - \frac{I(C_i \geq t)\, e^{\boldsymbol{Z}_i'\boldsymbol{\tau}}}{n\, S^{(0)}(t, \boldsymbol{\tau})}\, d\bar{N}(t) \right] \, ,$$

$i = 1, ..., n$. Sun and Wei (2000) show that for large n, the distribution of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ can be approximated by a normal distribution with mean 0 and covariance matrix $D(\tilde{\boldsymbol{\theta}})\,\tilde{\Gamma}\,D'(\tilde{\boldsymbol{\theta}})$, where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$,

$$D(\boldsymbol{\theta}) = \left(A^{-1}(\boldsymbol{\theta}), -B^{-1}(\gamma), -A^{-1}(\boldsymbol{\theta})\, P(\boldsymbol{\theta})\, G^{-1}(\boldsymbol{\tau}) \right)$$

and

$$\tilde{\Gamma} = \sum_{i=1}^{n} \begin{pmatrix} \tilde{a}_i(\tilde{\boldsymbol{\theta}}) + \tilde{b}_i(\tilde{\boldsymbol{\theta}}) \\ \tilde{d}_i(\tilde{\gamma}) \\ \tilde{d}_i(\tilde{\boldsymbol{\tau}}) \end{pmatrix} \left(\tilde{a}_i'(\tilde{\boldsymbol{\theta}}) + \tilde{b}_i'(\tilde{\boldsymbol{\theta}}) \,,\, \tilde{d}_i'(\tilde{\gamma}) \,,\, \tilde{d}_i(\tilde{\boldsymbol{\tau}})' \right) .$$

Let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\tau}_0$ denote the true values of $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$, respectively. Then it can be easily shown (Lawless and Nadeau, 1995; Sun and Wei, 2000) that for large n, the distributions of $\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ and $\tilde{\boldsymbol{\tau}} - \boldsymbol{\tau}_0$ can be approximated by normal distributions with mean zero and covariance matrices

$$B^{-1}(\tilde{\gamma}) \left[\sum_{i=1}^{n} \tilde{d}(\tilde{\gamma})\, \tilde{d}'(\tilde{\gamma}) \right] B^{-1}(\tilde{\gamma})$$

and

$$G^{-1}(\tilde{\boldsymbol{\tau}}) \left[\sum_{i=1}^{n} \tilde{d}(\tilde{\boldsymbol{\tau}})\, \tilde{d}'(\tilde{\boldsymbol{\tau}}) \right] G^{-1}(\tilde{\boldsymbol{\tau}}) ,$$

respectively.

9.4.2.3 Estimation with $\boldsymbol{\tau} = 0$

Sometimes it may be reasonable to assume that the C_i's are independent and identically distributed, that is, $\boldsymbol{\tau} = 0$. In this case, an estimation procedure similar to, but simpler than the one given above, can be developed. To see this, note that under current situation, $S_i(t)$ in (9.11) is independent of i. This suggests an unbiased estimating function

$$U_1(\boldsymbol{\beta}, \gamma) = \sum_{i=1}^{n} \boldsymbol{Z}_i \, \exp[-\boldsymbol{Z}_i' \, (\boldsymbol{\beta} + \gamma)] \int N_i(t)\, d\tilde{N}_i(t)$$

for estimation of $\boldsymbol{\beta}$ with given $\boldsymbol{\gamma}$. Let $\tilde{\boldsymbol{\beta}}_1$ denote the estimator of $\boldsymbol{\beta}$ given by the solution to $U_1(\boldsymbol{\beta}, \tilde{\gamma}) = 0$. It can be easily shown that $\tilde{\boldsymbol{\beta}}_1$ is consistent and unique (Sun and Wei, 2000). Furthermore, for large n, one can approximate the distribution of $\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0$ by a normal distribution with mean zero and covariance matrix

$$\left(A_1^{-1}(\tilde{\boldsymbol{\beta}}_1 + \tilde{\gamma}), -B^{-1}(\tilde{\gamma}) \right) \tilde{\Gamma}_1 \left(A_1^{-1}(\tilde{\boldsymbol{\beta}}_1 + \tilde{\gamma}), -B^{-1}(\tilde{\gamma}) \right)' ,$$

where $A_1(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and

$$\tilde{\Gamma}_1 = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i' e_i^{*2} e_i^2 & \sum_{i=1}^{n} \boldsymbol{Z}_i \tilde{d}_i'(\tilde{\gamma}) e_i^* e_i \\ \sum_{i=1}^{n} \tilde{d}_i(\tilde{\gamma}) \boldsymbol{Z}_i' e_i^* e_i & \sum_{i=1}^{n} \tilde{d}_i(\tilde{\gamma}) \tilde{d}_i'(\tilde{\gamma}) \end{pmatrix}$$

with $e_i = \int N_i(t) d\tilde{N}_i(t)$ and $e_i^* = \exp\{-\boldsymbol{Z}_i' \, (\tilde{\boldsymbol{\beta}}_1 + \tilde{\gamma})\}$, $i = 1, ..., n$.

9.4.3 An Example

Consider data set V of Appendix A, a set of panel count data reproduced from Sun and Wei (2000) and arising from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Byar et al., 1977; Byar, 1980). The study consists of the patients who had superficial bladder tumors when they entered the study, and these tumors were removed transurethrally. Many patients had multiple recurrences of tumors during the study, and these recurrent tumors were also removed transurethrally at the patient's clinic visits. The observed information includes clinical visit times and the numbers of recurrent tumors that occurred between the visits for each patient. In addition, for each patient, two potentially important baseline covariates were recorded, and they are the number of initial tumors and the size of the largest initial tumor.

Data set V of Appendix A consists of 85 patients who were randomly allocated at the beginning of the study to one of two treatments, placebo (47) and thiotepa (38). The original study involves a third treatment pyridoxine and is not considered here. The unit for observation times is a month with the largest observation time being 53 months. One of the study objectives was to make inferences about the effects of the treatments and baseline covariates on tumor recurrence rates.

To estimate the treatment and covariate effects, for patient i, define Z_{i1} to be the initial number of tumors observed at the beginning of the study, Z_{i2} the size of the largest initial tumor, and $Z_{i3} = 1$ if the patient is in the thiotepa group and 0 otherwise, $i = 1, ..., 85$. Also define $N_i(t)$ and $\tilde{N}_i(t)$ to be the total number of bladder tumors that had occurred and the number of clinical visits up to month t for patient i, respectively. First we apply the non-homogeneous Poisson approach, which gives $\hat{\boldsymbol{\beta}} = (0.2831, -0.0515, -1.3574)'$. For variance estimation, using the simple bootstrap procedure with 200 resamples of the observed data, we obtain the estimated standard errors 0.0832, 0.1007, and 0.3695 for $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, respectively.

Application of the estimating equation approach to the bladder tumor data yields $\tilde{\boldsymbol{\beta}} = (0.6620, -0.1229, -2.0249)'$ with estimated standard errors $(0.2133, 0.2035, 0.4500)'$. The non-homogeneous Poisson and estimating equation approaches give similar results. Both suggest that the recurrent rate of the tumor was significantly related to the number of initial tumors and that the thiotepa treatment significantly reduced the recurrent rate. On the other hand, the size of the largest initial tumor seems to have no significant effect on the recurrence of the bladder tumor. It is noted that the estimating equation approach indicates a more significant effect of the thiotepa treatment than the non-homogeneous Poisson approach. A possible reason for this is that the former takes into account the possible treatment effect on observation and follow-up times.

For the data set, the application of the estimating equation approach also gives $\tilde{\boldsymbol{\gamma}} = (-0.0094, 0.0392, 0.5084)'$ and $\tilde{\boldsymbol{\tau}} = (-0.0083, -0.1186, 0.1571)'$

with their estimated standard errors equal to $(0.0350, 0.0325, 0.0674)'$ and $(0.3955, 0.3352,\ 0.6904)'$, respectively. These results indicate that the patients in the thiotepa group visited the clinics more often than those in the placebo group, and the follow-up or censoring time did not seem to depend on the treatment or covariates. Note that the large number of visits by the patients in the thiotepa group could be because the thiotepa was installed into the patient's bladder and thus they needed more visits.

From the results obtained above, it seems reasonable to assume that $\boldsymbol{\tau} = 0$. Under this assumption, the estimating equation approach gives $\tilde{\boldsymbol{\beta}}_1 = (0.6604, -0.1230,\ -1.9712)'$ with estimated standard errors being $(0.2247, 0.2043, 0.4423)'$. These results are close to those given without assuming $\boldsymbol{\tau} = 0$.

9.5 Bibliography, Discussion, and Remarks

The analysis of panel count data is still a relatively new and not-well developed field, and there remain many open problems. One of the pioneering works in the field was given by Kalbfleisch and Lawless (1985), which dealt with panel count data arising from finite state Markov model. The other early work in the field include Breslow (1984), Gaver and O'Muircheartaigh (1987), Hinde (1982), Thall (1988, 1989), Thall and Lachin (1988), and Thall and Vail (1990). Most of these references mainly focus on parametric approaches for the analysis of panel count data. In terms of nonparametric and semiparametric approaches for panel count data, as discussed above, Sun and Kalbfleisch (1995), Weller and Zhang (2000), and Zhang and Jamshidian (2003) considered the one-sample nonparametric estimation problem. The authors who studied the nonparametric comparison problem include Sun and Fang (2003), Sun and Kalbfleisch (1993), and Sun and Rai (2001). The first two references dealt with comparison of the mean functions of two counting processes, while the third one discussed comparison of the intensity processes of several counting processes. In particular, Sun and Rai (2001) gave some comparison procedure for the situation where the ratios of the numbers of subjects between different groups are approximately constant over the whole study period.

The authors who studied regression analysis of panel count data include Cheng and Wei (2000), Hu et al. (2003), Ishwaran and James (2004), Lawless and Zhan (1998), Staniswalls et al. (1997), and Sun and Matthews (1997) in addition to Sun and Wei (2000), Wellner et al. (2004), and Zhang (2002) discussed above. In particular, Cheng and Wei (2000) and Hu et al. (2003) considered the estimating equation approach for situations where $\boldsymbol{\gamma} = \boldsymbol{\tau} = 0$ or $\boldsymbol{\tau} = 0$, respectively. They derived their estimating equations based on the counting processes

$$\int_0^t N_i(s)\, d\tilde{N}_i(s)\,,\ t \geq 0\,,\ i = 1, ..., n\,.$$

Note that for these processes, recurrent event data are available. For the same problem, Ishwaran and James (2004) employed the approach that models the intensity process of underlying counting processes instead of the mean function. In contrast, Lawless and Zhan (1998) and Staniswalls et al. (1997) suggested to analyze the data through modeling the rate function. The former employed a piecewise constant rate function assumption, whereas the latter used the Poisson assumption and smoothing techniques. In Sun and Matthews (1997), the focus was on the case where real observation times can be modeled as variations of prespecified observation times.

A couple of other issues about panel count data were also considered in the literature. For example, Chen et al. (2005) discussed the analysis of multivariate panel count data using a marginal mixed Poisson process approach by assuming that the baseline intensity function is piecewise constant. Sinha and Maiti (2004) considered the analysis of panel count data when the censoring time may be related to the counting process of interest for the situation where all study subjects have the same and fixed observation times.

In the preceding sections, the focus is mainly on inference about the mean function of underlying counting processes for panel count data. As mentioned before, given the structure of panel count data and the amount of observed information, it is much more convenient to deal with the mean function rather than the intensity process or rate function. On the other hand, sometimes one may want to directly model the intensity process or rate function (Ishwaran and James, 2004; Lawless and Zhan, 1998; Staniswalls et al., 1997). For this purpose, however, one usually has to make certain assumptions about the shape of the intensity process or rate function and/or the observation process in order to perform nonparametric or semiparametric analysis (Sun and Rai, 2001; Sun and Matthews, 1997).

The methods presented in this chapter mainly focus on panel count data in which observation and censoring times differ from subject to subject. For the situation where observation times or intervals are the same for all subjects, the data can be regarded as multivariate data, and any method that accommodates multivariate positive integer-valued response variables can be used for the analysis. This holds even though subjects may miss some intermediate observations and/or drop out of the study early. In this case, the resulting data can be seen as multivariate data with missing values.

Besides those discussed above, a general approach to the analysis of panel count data is to apply the methods developed for general longitudinal data analysis. For example, one could regard $\{n_{i,1}, ..., n_{i,m_i}\}$ or $\{n_{i,1}, n_{i,2} - n_{i,1}, ..., n_{i,m_i} - n_{i,m_i-1}\}$ as repeated measurements on subject i. One major disadvantage of this approach is that it is usually difficult to make use of the underlying monotonic property of the mean and thus is less efficient.

A basic assumption underlying all the methods discussed in this chapter is that the observation process, or the observation and censoring times, are independent of the counting process of interest completely or given covariates. In practice, this may not be true. There exists a great deal of research in the literature concerning the analysis of longitudinal data in the presence of a related or dependent censoring time. Sometimes this censoring time is also referred to as drop-out time in the longitudinal data literature (Little, 1995; Wang and Taylor, 2001; Wulfsohn and Tsiatis, 1997). Some authors also discussed longitudinal data analysis when response and observation processes may be related (Lin and Scharfstein, 2004; Robins et al., 1995; Sun et al., 2005). Among others, Huang and Wang (2004) and Wang et al. (2001) considered the analysis of recurrent event data when the censoring time may be correlated with the underlying counting process of interest. However, except Sinha and Maiti (2004), there exists little work that studies panel count data when the underlying counting process of interest and observation process may depend on each other.

For a recurrent event, in addition to the occurrence times of the event, one may also be interested in the gap times, which are defined as the times between successive occurrences of the event (Chen et al., 2004; Huang and Chen, 2003; Sun, Park, and Sun, 2006; Zhao and Sun, 2006). Also, there may be several different, but related recurrent events that appear in the same study and need to be studied together (Chen et al., 2005). For their analyses based on panel count data, there exists relatively limited research in the literature.

10

Other Topics

10.1 Introduction

In this chapter, we discuss several important topics about interval-censored failure time data that can occur in practice but were not discussed in the previous chapters. These include goodness-of-fit (GOF) tests or regression diagnostics, regression analysis of failure time data with interval-censored covariates, Bayesian analysis of interval-censored failure time data, and informative interval censoring.

In Section 10.2, we start with discussion on issues related to regression diagnostics. For regression analysis of interval-censored data, as in general regression analysis, several questions need to be asked. Among them, the most important one is perhaps the adequacy or appropriateness of a regression model selected to fit the observed interval-censored data. In other words, one needs to assess if the model fits the data well or if there is another model that provides a better fit. The same question could be asked for a parametric model. Furthermore, one may ask if the functional form of the covariates in the regression model is appropriate or the best. Also, one may want to identify possible outliers. To address these questions, we present in Section 10.2 several simple, intuitive approaches that are direct generalizations of the methods developed for right-censored failure time data.

In the discussion about regression analysis in the preceding chapters, it was assumed that covariates can be observed exactly. In reality, however, the observation on covariates could suffer interval censoring as well as the survival time of interest. Section 10.3 presents approaches to some inference problems in the presence of interval-censored covariates. In Section 10.4, we consider Bayesian analysis of interval-censored failure time data, which is often performed when there exists prior information about the survival time of interest. Several issues concerning nonparametric and semiparametric Bayesian analyses are discussed.

Informative interval censoring is the topic of Section 10.5. In all of the preceding chapters, we have assumed that the interval censoring mechanism is independent of the survival time of interest. It is well-known that this may not always hold in practice, and informative interval censoring may occur if the independence assumption is violated. In Section 10.5, interval censoring mechanisms are discussed and some techniques are described that can be used for the analysis when informative interval censoring is present. Section 10.6, the last section of the book, considers computational aspects of the analysis of interval-censored data and several additional practical topics related to interval-censored failure time data.

10.2 Regression Diagnostics

It is well-known that in general, regression diagnostics should be conducted in each regression analysis unless, for example, the same regression problem and model have been considered before. For complete or right-censored failure time data, a number of graphical or quantitative methods have been developed for regression diagnostics. Among others, readers are referred to books by Collett (1994), Klein and Moeschberger (2003), and Lawless (2003). For interval-censored failure time data, however, there are not many approaches available for regression diagnostics, and most of the existing methods are graphical and intuitive with finite sample and asymptotic properties unknown.

First, we consider a parametric analysis of interval-censored failure time data in Section 10.2.1. In this situation, the most fundamental and commonly asked question in terms of diagnostics concerns the appropriateness of an assumed parametric model. To answer this question, one needs to perform a GOF test that assesses the overall fit of the model to a given data set. Section 10.2.1 gives some simple GOF test procedures. Section 10.2.2 deals with regression diagnostic procedures that can be applied to regression analysis of interval-censored data using the PH model. For this model, several graphical model-checking approaches are investigated including those based on residuals. In Section 10.2.3, for interval-censored data with the additive hazards model, some simple residual-based procedures for regression diagnostics are briefly discussed.

10.2.1 Parametric Regression Analysis

Consider a survival study that consists of n independent subjects for which each subject gives rise to a variable T_i, the survival time of interest, $i = 1, ..., n$. It is assumed that the T_i's follow a parametric model with density, cumulative distribution, and survival functions $f(t)$, $F(t)$, and $S(t)$, respectively. Also, we assume that for the T_i's, only interval-censored data are available and they have the form

$$\{ (L_i, R_i], ; i = 1, ..., n \} ,$$

where as before, $(L_i, R_i]$ denotes the interval to which T_i is observed to belong, $i = 1, ..., n$.

Suppose that we are interested in testing the null hypothesis

$$H_0^1 \,:\, F(t) \,=\, F_0(t)\,,$$

where $F_0(t)$ is a completely known cumulative distribution function. Let $0 = s_0 < s_1 < ... < s_m$ denote the ordered distinct time points of all left and right end time points L_i's and R_i's. Also let $\hat{F}_n$ denote the NPMLE of F, which can be obtained by the algorithms given in Section 3.4. Then to test H_0, it is apparent that a simple graphical approach is to plot $\hat{F}_n$ and F_0 together and to check for any major discrepancy between them.

To derive a quantitative GOF procedure, define $\alpha_{ij} \,=\, I((s_{j-1}, s_j] \subseteq (L_i, R_i])$ and for each j, let O_j and E_j denote the observed and expected numbers of subjects who fail within $(s_{j-1}, s_j]$, $j \,=\, 1, ..., m$, $i \,=\, 1, ..., n$. We note that for complete data or more generally, if all the O_j's are known, a common approach for testing H_0 is to apply the Pearson χ^2 statistic defined as

$$X_n^2 \,=\, \sum_{j=1}^{m} \frac{(O_j \,-\, E_j)^2}{E_j}\,. \tag{10.1}$$

Of course, the O_j's are unknown for interval-censored data, while the E_j's can be easily calculated as

$$E_j \,=\, n\,[\,F_0(s_j) \,-\, F_0(s_{j-1})\,]$$

under H_0, $j \,=\, 1, ..., m$. To use X_n^2, different methods can be used to estimate the O_j's. By using the same idea as for the calculation of the E_j's and treating $\hat{F}_n$ as the true distribution, natural estimates are

$$\hat{O}_j \,=\, n\,[\,\hat{F}_n(s_j) \,-\, \hat{F}_n(s_{j-1})\,] \tag{10.2}$$

or alternatively

$$\hat{O}_j \,=\, \sum_{i=1}^{n} \frac{\alpha_{ij}\,[\hat{F}_n(s_j) \,-\, \hat{F}_n(s_{j-1})]}{\hat{F}_n(R_i) \,-\, \hat{F}_n(L_i)}\,, \tag{10.3}$$

$j \,=\, 1, ..., m$.

Given estimates of the O_j's, one needs to know the distribution of the statistic X_n^2. Unlike the case of complete or right-censored data, the asymptotic distribution of X_n^2 has not been obtained for interval-censored data. A major reason is that, as discussed in Section 3.6, the convergence rate of $\hat{F}_n$ depends on the behavior of the observation process generating censoring intervals and is slower than $n^{1/2}$ unless there exists a large portion of uncensored observations. Using simulation, Babineau (2005) investigated the finite sample null distribution of X_n^2 with the $\hat{O}_j$'s given by (10.3) as well as its power and sensitivity. Other similar statistics for testing H_0 including the likelihood ratio test statistics were also studied in Babineau (2005) .

A statistic similar to X_n^2 with the $\hat{O}_j$'s given by (10.2) is the Cramer-von Mises statistic given by

$$U_n = n \int_0^\infty [\hat{F}_n(t) - F_0(t)]^2 \, dF_0(t) \, .$$

Ren (2003) argued that U_n is not suitable for testing H_0 because $\hat{F}_n$ does not converge uniformly and thus U_n is not stable under H_0 as $n \to \infty$. She proposed to test H_0 with the statistic

$$U_n^* = m \int_0^\infty [\hat{F}_{nm}^*(t) - F_0(t)]^2 \, dF_0(t)$$

and established its asymptotic distribution. In the statistic above, $\hat{F}_{nm}^*$ denotes the empirical distribution function of the so-called leveraged bootstrap sample of size m.

A hypothesis that is more general and practical than H_0 is

$$H_0^2 \, : \, F(t) = F_0(t, \theta) \, ,$$

where $F_0(t, \theta)$ is a cumulative distribution function known up to parameter θ. Let $\hat{\theta}$ denote the maximum likelihood estimator of θ based on the observed interval-censored data. To test H_0^2, one can apply the statistic X_n^2 with the O_j's and E_j's estimated by (10.3) and

$$\hat{E}_j = n \, [\, F_0(s_j, \hat{\theta}) - F_0(s_{j-1}, \hat{\theta}) \,] \, ,$$

respectively, or the statistic

$$U_n^*(\hat{\theta}) = m \int_0^\infty [\hat{F}_{nm}^*(t) - F_0(t, \hat{\theta})]^2 \, dF_0(t, \hat{\theta})$$

(Babineau, 2005; Ren, 2003).

10.2.2 Analysis with the Proportional Hazards Model

In this subsection, we discuss the same problem as in Section 6.2, but focus on regression diagnostics. Specifically, consider a survival study that consists of n independent subjects and gives rise to interval-censored data

$$\{ \, (L_i, R_i], \boldsymbol{Z}_i \, ; \, i = 1, ..., n \, \}$$

for the survival times T_i's of interest. Here $(L_i, R_i]$ denotes the interval within which the survival event for the ith subject is observed to occur, and $\boldsymbol{Z}_i$ represents the vector of covariates from subject i, $i = 1, ..., n$. Let $S(t; \boldsymbol{Z})$ denote the survival function for a subject with covariate $\boldsymbol{Z}$ and suppose that it is specified by the PH model (1.4). Then the log likelihood function is given by

$$l(\boldsymbol{\beta}, S_0) = \sum_{i=1}^n \log \left\{ \, [S_0(L_i)]^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} - [S_0(R_i)]^{\exp(\boldsymbol{Z}_i' \boldsymbol{\beta})} \, \right\} \tag{10.4}$$

in terms of regression parameters $\boldsymbol{\beta}$ and the baseline survival function $S_0(t)$. As before, we use $\hat{\boldsymbol{\beta}}_n$ and $\hat{S}_n(t)$ to denote the maximum likelihood estimators of $\boldsymbol{\beta}$ and S_0, which are defined in Section 6.2. Also define $\hat{S}(t;\boldsymbol{Z}) = [\hat{S}_n(t)]^{\exp(\boldsymbol{Z}'\hat{\boldsymbol{\beta}})}$.

For the case of right-censored failure time data, various types of residuals have been proposed for assessing the appropriateness of the PH model analysis. These include Cox-Snell residuals (Cox and Snell, 1968), deviance residuals (Therneau et al., 1990), martingale residuals (Barlow and Prentice, 1988; Lagakos, 1980), and Schoenfeld residuals (Schoenfeld, 1982). In the following, we discuss several generalizations of these residuals and their uses in the analysis of interval-censored data.

The Cox-Snell residual was proposed by Cox and Snell (1968) to assess the overall fit of the PH model to right-censored failure time data and is based on the fact that the variable $-\log S(T;\boldsymbol{Z})$ has an exponential distribution with hazard rate one. Based on their work, Farrington (2000) suggests that for interval-censored data, we can define the Cox-Snell residual interval as $(-\log \hat{S}(L_i;\boldsymbol{Z}_i), -\log \hat{S}(R_i;\boldsymbol{Z}_i)]$ for subject i, $i = 1,...,n$. To use them for assessing GOF, as in the case of right-censored data, one can estimate the cumulative hazard function corresponding to these residual intervals and plot the estimated function against the distinct values of $\{-\log \hat{S}(L_i;\boldsymbol{Z}_i)\,,\, -\log \hat{S}(R_i;\boldsymbol{Z}_i)\}_{i=1}^n$. If the PH model provides a reasonable fit, the plot should approximately give a straight line of unit slope through the origin. Through some examples of parametric data analyses, Farrington (2000) points out that this approach may not be very sensitive to the PH model assumption and thus not provide a good diagnostic tool. More research on these residuals and this approach is needed.

Among all others, the martingale residual, also sometimes referred to as the Lagakos residual, may be more commonly used. For right-censored failure time data, the martingale residuals can be defined as the scores given by the score function of the regression parameter $\boldsymbol{\beta}$. These residuals have sample mean zero. For the current situation, it can be easily shown from (10.4) that the score function of $\boldsymbol{\beta}$ has the form

$$\sum_{i=1}^n \boldsymbol{Z}_i \, \frac{S(L_i;\boldsymbol{Z}_i)\,\log S(L_i;\boldsymbol{Z}_i) \,-\, S(R_i;\boldsymbol{Z}_i)\,\log S(R_i;\boldsymbol{Z}_i)}{S(L_i;\boldsymbol{Z}_i) \,-\, S(R_i;\boldsymbol{Z}_i)} \,.$$

This suggests that when fitting interval-censored data to the PH model (1.4), one can define the martingale residual as

$$M_i \,=\, \frac{S(L_i;\boldsymbol{Z}_i)\,\log S(L_i;\boldsymbol{Z}_i) \,-\, S(R_i;\boldsymbol{Z}_i)\,\log S(R_i;\boldsymbol{Z}_i)}{S(L_i;\boldsymbol{Z}_i) \,-\, S(R_i;\boldsymbol{Z}_i)}$$

with $S(t;\boldsymbol{Z}_i)$ replaced by $\hat{S}(t;\boldsymbol{Z}_i)$ and expect that they have mean zero for large n, $i = 1,...,n$. Farrington (2000) considered the use of the M_i's for parametric analyses.

The martingale residuals can be plotted in several different ways as in the case of right-censored data. One is to plot them against continuous covariates to examine the functional form of covariates in the PH model. Also one could examine the plot of the residuals against covariates not included in the model to assess the need of including these covariates in the model, especially for categorical covariates. Another plot is the residuals against observation number for identifying observations that are outliers. The martingale residuals have also been used to construct GOF test statistics for the PH model for right-censored data (Lin et al., 1993), but it is unclear how to do the same for interval-censored data.

To identify outlier observations or assess the influence of individual observations, deviance residuals are more commonly used. As with right-censored data, for interval-censored data, we can define the deviance as $D = -2\,l(\boldsymbol{\beta}, S_0)$ with parameters replaced by their maximum likelihood estimators. Also for each i, define D_i to be the deviance D with the ith observation removed from the data set. Let $\hat{\beta}_j$ denote the maximum likelihood estimator of the regression parameter representing the effect of the jth covariate and $\hat{\beta}_{j(i)}$ the corresponding estimator with the ith observation removed. Younes and Lachin (1997) define

$$DR_i = D - D_i$$

and

$$DR_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{sd(\hat{\beta}_j)}$$

as the deviance residuals for observation i and the jth covariate with respect to observation i, respectively, where $sd(\hat{\beta}_j)$ denotes the estimated standard deviation of $\hat{\beta}_j$. For their applications, they suggest plotting them against the observation number to detect any observations that are not consistent with the other observations. One advantage of the deviance residual over the martingale residual is that the latter ranges between $-\infty$ and 1 and is skewed, while the former can be expected to have a pattern close to that given by a normal variable. As with other residuals, however, one should use the deviance residual only as exploratory graphical tools because their asymptotic and finite sample distributions are unknown.

A key assumption about the PH model is that of proportionality of the hazard rates for subjects with distinct values of a covariate. To check this, a graphical approach is the simplest and most straightforward. For example, for a covariate taking only a finite number of values, one can estimate survival functions for subjects with the same covariate values separately as well as under model (1.4), respectively, and plot them together as in Section 4.5. An alternative is to plot the log minus log of the survival functions estimated separately or the log of the separately estimated cumulative hazard functions. Specifically, suppose that the covariate Z_1 takes k different values. Based on the observed data for the subjects with Z_1 taking its jth value, let $\hat{\Lambda}_j(t) = -\log \hat{S}_j(t)$ denote the estimated cumulative hazard or minus the log survival function given in Section 3.4 , $j = 1, ..., k$. Then one could plot $\log \hat{\Lambda}_1(t), ..., \log \hat{\Lambda}_k(t)$ together versus t, which should be approximately parallel to each other if the PH model is correct. Alternatively, we can plot $\log \hat{\Lambda}_2(t) - \log \hat{\Lambda}_1(t), ..., \log \hat{\Lambda}_k(t) - \log \hat{\Lambda}_1(t)$ versus t, which are supposed to be roughly constant under the PH model. For a continuous covariate, one could group its values into k different intervals and then apply the graphical tools discussed above.

10.2.3 Analysis with the Additive Hazards Model

Now we consider regression diagnostics for regression analysis of interval-censored failure time data using the additive hazards model (1.7). In this case, even for right-censored data, there do not exist many procedures for regression diagnostics. For a GOF test of model (1.7) with interval-censored data, an easy and natural graphical tool is to plot the log of the estimated survival function or the estimated cumulative hazard function for the values of a categorical covariate. Specifically, suppose that the covariate Z_1 takes k different values and let the $\hat{\Lambda}_j(t)$'s be defined as in the previous subsection. To check the additive relationship between the hazard functions for subjects with different values of Z_1, we can plot $\hat{\Lambda}_1(t), ..., \hat{\Lambda}_k(t)$ or $\hat{\Lambda}_2(t) - \hat{\Lambda}_1(t), ..., \hat{\Lambda}_k(t) - \hat{\Lambda}_1(t)$ together versus t. If model (1.7) is correct, the former plot should give approximately parallel curves, while the latter plot should be roughly displays of some constants. As for the PH model, if Z_1 is continuous, we can group it to a categorical variable.

Suppose that the survival study of interest gives rise to current status data. In this case, using the notation defined in Section 5.4, the martingale residual can be defined as $\hat{M}_i(t) = M_i(t)$ given in (5.11) with $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\Lambda_0^*(t)$ replaced by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ defined there and

$$\hat{\Lambda}_0^*(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{n^{-1}\, S^{(0)}(s; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})} \, dN_i(s) \,,$$

respectively. Ghosh (2003) first studied this martingale residual and suggested plotting the $\hat{M}_i(\infty)$'s against the values $\{ Z_{ji} \}$ of a covariate Z_j to assess the functional form of the covariate. He further constructed the formal test statistic

$$W_j(z) = n^{1/2} \sum_i I(Z_{ji} \leq z) \, \hat{M}_i(\infty)$$

and showed that under model (1.7), the distribution of $W_j(z)$ can be approximated by that of a Gaussian process $\hat{W}_j$ with mean zero. Thus for checking the functional form of Z_j, one can plot $W_j(z)$ along with some realizations of $\hat{W}_j(z)$ versus z. Under model (1.7), the curve given by $W_j(z)$ should be like one of the others.

10.3 Regression Analysis with Interval-censored Covariates

As remarked in Section 1.3.5, in practice, interval censoring can occur on observations on covariates as well as on those on the survival time of interest. Goggins et al. (1999b) give an example of failure time data with an interval-censored covariate arising from the ACTG 181, the same AIDS clinical trial producing the data presented in data set I of Appendix A and discussed in Section 1.2.4. In the study, an objective of interest is to model the possible relationship between the status of CMV shedding in the blood and urine and the onset of active CMV end-organ disease. For this, they proposed to fit the PH model (1.4) treating the onset time of CMV disease as the survival time of interest and the status of CMV shedding, presence or absence, as a binary time-dependent covariate. For the onset time, the study yielded right-censored data. For the binary covariate, however, only interval-censored data are available because as discussed in Section 1.2.4, only interval-censored data are available for the time to CMV shedding. A similar example, also arising from an AIDS clinical trail, can be found in Gómez et al. (2003).

Two types of situations regarding interval-censored covariates are considered. First we deal with the situation in which one observes right-censored failure time data on the survival time of interest. The situation where doubly censored data are available is then discussed. For the latter, both covariates and survival time suffer from interval censoring. Some regression models and inference approaches are discussed for each case, and the section is concluded with remarks on some related issues.

10.3.1 Analysis with Right-censored Data

This subsection concerns a survival study that consists of n independent subjects and gives rise to right-censored failure time data

$$\{Q_i = \min\{T_i, C_i\}\,,\, \delta_i = I(Q_i = T_i)\,,\, (Z_{Li}, Z_{Ri}]\,,\, \boldsymbol{W}_i\,;\, i = 1, ..., n\}$$

for the T_i's, the survival times of interest. In the expression above, for subject i, C_i denotes a right censoring variable assumed to be independent of T_i, $(Z_{Li}, Z_{Ri}]$ represents an interval to which a scalar covariate Z_i is observed to belong, and $\boldsymbol{W}_i$ denote all other covariates that are assumed to be known or observed exactly. In other words, we observe right-censored data for the T_i's and exact data for the W_i's, but only interval-censored data for the covariate Z_i's. We discuss below the situation where there exist more than one covariate whose observations are interval-censored. The goal is to make inferences about effects of the covariates $\boldsymbol{W}_i$ and Z_i on T_i.

Suppose that the effects of covariates can be described by the PH model

$$\lambda(t; \boldsymbol{W}_i, Z_i) = \lambda_0(t) \exp(\boldsymbol{W}_i' \boldsymbol{\beta} + Z_i \gamma) \tag{10.5}$$

in terms of the hazard function of T_i given $\boldsymbol{W}_i$ and Z_i. In the model above, as before, λ_0 is an unknown baseline hazard function and $\boldsymbol{\beta}$ and γ represent the effects of $\boldsymbol{W}_i$ and Z_i on T_i, respectively. Define $\boldsymbol{Z} = (Z_1, ..., Z_n)'$ and $Y_i(t) = I(Q_i \geq t)$, $i = 1, ..., n$. Let G denote the cumulative distribution function of the Z_i's. For inference about covariate effects, if $\boldsymbol{Z}$ is known, as discussed before, we can simply apply the partial likelihood function

$$L(\boldsymbol{\beta}, \gamma; \boldsymbol{Z}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{W}_i' \boldsymbol{\beta} + Z_i \gamma)}{\sum_{j=1}^{n} Y_j(Q_i) \exp(\boldsymbol{W}_j' \boldsymbol{\beta} + Z_j \gamma)} \right]^{\delta_i} . \qquad (10.6)$$

By treating the function above as a conditional likelihood given $\boldsymbol{Z}$, one can write down the full likelihood function given the $\boldsymbol{W}_i$'s as

$$L(\boldsymbol{\beta}, \gamma, G) = L(\boldsymbol{\beta}, \gamma; \boldsymbol{Z}) \prod_{i=1}^{n} dG(Z_i) \qquad (10.7)$$

because as assumed, all covariates are external (Kalbfleisch and Prentice, 2002). Of course, we cannot use either the partial likelihood (10.6) or the full likelihood (10.7) because the Z_i's are unknown. On the other hand, one could apply them to develop some EM algorithms for estimation of $\boldsymbol{\beta}$ and γ.

Goggins et al. (1999b) first considered this EM estimation approach for the situation where $\boldsymbol{W}_i = 0$ and Z_i is defined as $Z_i(t) = I(S_i \leq t)$, indicating by one if an event with occurrence time S_i has occurred. In their application, S_i denoted the time to onset of CMV shedding in blood and urine, which we discussed at the beginning of this section. We remark that although Z_i is time-dependent, it still fits the general situation discussed here because it is a binary variable. For the current situation, $(Z_{Li}, Z_{Ri}]$ represents the observed interval within which Z_i switches from 0 to 1 or to which S_i belongs. For estimation of γ, Goggins et al. (1999b) developed a Monte Carlo EM algorithm that involves imputing the values of the S_i's.

Instead of the PH model (10.5), one can apply the following log-linear model

$$T_i^* = \log T_i = \boldsymbol{W}_i' \boldsymbol{\beta} + Z_i \gamma + \epsilon_i \qquad (10.8)$$

for the T_i's, where the ϵ_i's represent random errors. Gómez et al. (2003) studied this model for situations where T_i^* is a general random variable, but observed exactly, and the distribution of the ϵ_i's belongs to a parametric family. Assuming that the distribution of the Z_i's has finite support, they developed a two-step conditional algorithm that iterates estimation of the distribution of the Z_i's and the other parameters. Topp and Gómez (2004) discussed the same problem focusing on residual analysis.

In theory, one can easily generalize the inference approaches discussed above to situations in which more than one covariate suffer interval censoring. However, implementation could be much more complicated. For example, consider a PH model analysis in which there exist two covariates Z_1 and Z_2 whose observations are interval-censored. In this case, to develop an EM algorithm, instead of $L(\boldsymbol{\beta}, \gamma, G)$ given in (10.7), one has to work with the likelihood

$$L(\boldsymbol{\beta}, \gamma; \boldsymbol{Z})\, L_{(Z_1, Z_2)} \,.$$

Here $L_{(Z_1,Z_2)}$ denotes the marginal likelihood given by samples of Z_1 and Z_2 and depends on the joint distribution of Z_1 and Z_2.

There are many open questions for the situation discussed here. For the Monte Carlo EM algorithm discussed above, no asymptotic justification is available yet for the derived parameter estimates. Another question of interest for model (10.5) concerns estimation of the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$. For model (10.8), the inference approach given in Gómez et al. (2003) applies only to exactly observed Y_i and requires that the distribution of the ϵ_i's is known up to some parameters. In reality, one might have data for which the accelerated failure time model (1.8) or some other semiparametric model is more appropriate than model (10.5) or (10.8) and some corresponding inference approaches are needed.

10.3.2 Analysis with Doubly Censored Data

Now we consider the same situation as that discussed in Section 8.3.2, but with one covariate whose observations are interval-censored. Let the X_i's, S_i's and T_i's be defined as in Section 8.3.2 and suppose that for the X_i's and S_i's, interval- and right-censored data are observed and given by

$$\{(L_i, R_i]\,,\, i = 1, ..., n\}\,,\ \{S_i^* = \min(S_i, C_i)\,,\, \delta_i = I(S_i^* = S_i)\,,\, i = 1, ..., n\}\,,$$

respectively, as before. Here the C_i's denote right censoring times with respect to the S_i's and are assumed to be independent of the S_i's as before. Also it is assumed that the X_i's and T_i's are independent. As in the previous subsection, for subject i, suppose that there is one interval-censored covariate Z_i with observation $(Z_{Li}, Z_{Ri}]$, i.e., $Z_i \in (Z_{Li}, Z_{Ri}]$, and let $\boldsymbol{W}_i$ denote all other covariates assumed to be known or observed exactly. Also suppose that the hazard function of T_i given $\boldsymbol{W}_i$ and Z_i is specified by the PH model (10.5).

For inferences about $\boldsymbol{\beta}$ and γ, let $\boldsymbol{X}$, $\hat{H}$, the $Y_i(t|X_i)$'s, $N_i(t|X_i)$'s and a_i's be defined as in Section 8.3.2. Furthermore, define

$$S^{(j)}(t; \boldsymbol{\beta}, \gamma | \boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t|X_i)\, \boldsymbol{Z}_i^j\, e^{\boldsymbol{W}_i' \boldsymbol{\beta} + Z_i \gamma}\,,$$

where $\boldsymbol{Z}_i = (\boldsymbol{W}_i', Z_i)'$, $\boldsymbol{Z}_i^0 = 1$, $\boldsymbol{Z}_i^1 = \boldsymbol{Z}_i$, $j = 0, 1$. Also, similar to $U_p(\boldsymbol{\beta}|\boldsymbol{X})$ given in (8.6), define

$$U_p(\boldsymbol{\beta}, \gamma | \boldsymbol{X}) = \int_0^\tau \sum_{i=1}^{n} \left[\boldsymbol{Z}_i - \frac{S^{(1)}(t; \boldsymbol{\beta}, \gamma | \boldsymbol{X})}{S^{(0)}(t; \boldsymbol{\beta}, \gamma | \boldsymbol{X})} \right] dN_i(t|X_i)\,,$$

where τ denotes the longest possible follow-up time as before. If the Z_i's are known exactly, following the discussion in Section 8.3.2, we can estimate $\boldsymbol{\beta}$ and γ by the solution to

$$U_p(\boldsymbol{\beta}, \gamma, \hat{H} | Z_i's) = \left(\prod_{i=1}^{n} a_i^{-1} \right) \int_{L_1}^{R_1} ... \int_{L_n}^{R_n} U_p(\boldsymbol{\beta}, \gamma \,|\, \boldsymbol{x}) \prod_{i=1}^{n} \left[d\hat{H}(x_i) \right] = 0 \,,$$

where $\boldsymbol{x} = (x_1, ..., x_n)$. This suggests that for the current situation, one can apply the same idea and employ the estimating equations

$$\left(\prod_{i=1}^{n} b_i^{-1} \right) \int_{Z_{L1}}^{Z_{R1}} ... \int_{Z_{Rn}}^{Z_{Rn}} U_p(\boldsymbol{\beta}, \gamma, \hat{H} \,|\, z_i's) \prod_{i=1}^{n} \left[d\hat{F}_Z(z_i) \right] = 0 \qquad (10.9)$$

for estimation of $\boldsymbol{\beta}$ and γ. In these equations, $\hat{F}_Z$ denotes the NPMLE of the cumulative distribution function of the Z_i's based on the interval-censored data on the Z_i's only and $b_i = \int_{Z_{Li}}^{Z_{Ri}} d\hat{F}_Z(z_i)$, $i = 1, ..., n$.

This estimation approach is a generalization of the partial likelihood approach for right-censored failure time data with known covariates and has the advantage that it does not involve the baseline hazard function. However, efficiency could be an issue and needs to be investigated. It was first proposed by Zhao et al. (2005) for the situation where $\boldsymbol{W}_i = 0$, i.e., Z_i is the only covariate for each study subject. Under some regularity conditions, they show that the resulting regression parameter estimate is consistent and its distribution can be asymptotically approximated by a normal distribution. It is straightforward to generalize these results to the current situation. One possible difficulty with the estimating equation (10.9) is that it may not be easy tov

ject. U3that

10.3.3 Some Remarks

In addition to the two situations discussed in the previous subsections, one can face interval-censored covariates when one observes case I or II interval-censored data on the survival time of interest. In this situation, inference about regression parameters is more difficult than in the two preceding situations and the ideas discussed above do not seem directly applicable. Instead, one may have to employ the full likelihood approach that involves the distributions of both survival variable and interval-censored covariates. Interval-censored covariates can also occur when, as in Chapter 7, there exist more than one survival variables of interest.

The analysis of right-censored failure time data with missing covariates is related to the topic of this section and has been investigated by many authors. Interval-censored covariates apparently differ from missing covariates because the former does provide some information about the covariates. For the latter, it is common to impute the missing covariates and this can be done for interval-censored covariates using the techniques discussed in Section 2.4 for interval-censored survival times. As with missing covariates, interval-censored covariates can also occur in contexts other than failure time data analysis. For instance, Chen and Cook (2003) discussed a situation where the response is a point process and the covariate is a marker process. For the response, complete or recurrent event data are observed, but the marker process is subject to interval censoring.

10.4 Bayesian Analysis

Using the notation defined before, consider a survival study that involves n independent subjects and gives rise to interval-censored data

$$\boldsymbol{O} \,=\, \{\, (L_i, R_i], \boldsymbol{Z}_i \,;\, i \,=\, 1, ..., n \,\}$$

for the survival time T_i of interest. Let $S(t; \boldsymbol{Z}, \boldsymbol{\theta})$ denote the survival function for a subject with covariates $\boldsymbol{Z}$ and suppose that it is known up to an unknown vector of parameters $\boldsymbol{\theta}$. Here the dimension of $\boldsymbol{\theta}$ may be finite or infinite. Then the likelihood function is proportional to

$$L(\boldsymbol{\theta} \,|\, \boldsymbol{O}) \,=\, \prod_{i=1}^{n} \left[\, S(L_i; \boldsymbol{Z}_i, \boldsymbol{\theta}) \,-\, S(R_i; \boldsymbol{Z}_i, \boldsymbol{\theta}) \,\right] \,,$$

and our goal is to make inference about $\boldsymbol{\theta}$.

For a Bayesian analysis, assume that $\boldsymbol{\theta}$ is random and follows a known distribution, say $\pi(\boldsymbol{\theta})$, which is called the prior distribution. Then one can make inference about $\boldsymbol{\theta}$ based on the distribution

$$\pi(\theta \,|\, \boldsymbol{O}) \,=\, \frac{L(\boldsymbol{\theta} \,|\, \boldsymbol{O}) \, \pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \, L(\boldsymbol{\theta} \,|\, \boldsymbol{O}) \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \,,$$

which is called the posterior distribution. In this expression, $\boldsymbol{\Theta}$ denotes the space of $\boldsymbol{\theta}$ and the probability formula used to derive $\pi(\theta \,|\, \boldsymbol{O})$ is Bayes's theorem. If the posterior distribution $\pi(\theta \,|\, \boldsymbol{O})$ has a closed form, one can draw inference about $\boldsymbol{\theta}$ from it. For example, one can estimate $\boldsymbol{\theta}$ by the posterior mean defined as the mean of $\boldsymbol{\theta}$ with respect to $\pi(\theta \,|\, \boldsymbol{O})$. In most applications, however, $\pi(\theta \,|\, \boldsymbol{O})$ does not have a closed form and in fact, the posterior distribution is often too complicated to be used directly. Then one may sample from it in order to make the inference.

In general, two basic tasks are involved in Bayesian analyses. One needs to select a prior distribution for $\boldsymbol{\theta}$ and obtain some computational methods to sample from the posterior distribution. For the case of right-censored failure time data, a number of authors have discussed these and other related issues. For instance, a number of sampling techniques based on Gibbs sampling and other Markov chain Monte Carlo sampling algorithms have been developed in the literature. For references, readers are referred to the review paper by Sinha and Dey (1997) and the recent book by Ibrahim et al. (2001), which provides a relatively complete and comprehensive coverage of Bayesian analysis approaches for failure time data. On the other hand, for interval-censored data, there exists much less research due to the difficulties introduced by such censoring. In the following, we briefly discuss two topics, nonparametric (Gómez et al., 2000) and semiparametric Bayesian approaches (Sinha et al., 1999) for the analysis of interval-censored data with the focus on the selection of prior distributions. Section 10.4.3 contains some general remarks.

10.4.1 Nonparametric Bayesian Approaches

In this subsection, we assume that all $\boldsymbol{Z}_i = 0$ or there exist no covariates and that $\boldsymbol{\theta}$ represents the cumulative distribution function F or the cumulative hazard function Λ of the T_i's. That is, the goal is to estimate $F(t)$ or $\Lambda(t)$ nonparametrically.

For nonparametric estimation of a cumulative distribution or hazard function, two prior processes are commonly used. One is the Dirichlet process prior introduced in Ferguson (1973; 1974) and the other is the beta process prior proposed in Hjort (1990). The former is usually used if $F(t)$ or $S(t) = 1 - F(t)$ is the target, while the latter is commonly applied if the focus is on $\Lambda(t)$.

For a set A on the real line, define $P(A) = \int_A dF(t)$, which is random if F is. The Dirichlet process prior assumes that for any partition of the real line denoted by $A_1, , ..., A_m$, the joint distribution of the random vector $\boldsymbol{Y} = (P(A_1), ..., P(A_m))$ is given by the Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha(A_1), ..., \alpha(A_m))$, where α is a probability measure. That is, $\boldsymbol{Y}$ has joint density function

$$f(y_1, ..., y_{m-1}) = \left[\frac{\Gamma(\sum_{j=1}^m \alpha_j)}{\prod_{j=1}^m \Gamma(\alpha_j)} \right] \left(\prod_{j=1}^{m-1} y_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{m-1} y_j \right)^{\alpha_m - 1} ,$$

$y_j \geq 0$, $j = 1, ..., m-1$, $\sum_{j=1}^{m-1} y_j \leq 1$, where Γ denotes the gamma function and $\alpha_j = \alpha(A_j)$, $j = 1, ..., m$. Assume that F is discrete and the corresponding survival variable T takes only finite values $s_1 < ... < s_m$ with $f_j = P(T = s_j)$, $j = 1, ..., m$. In this case, a common choice for the prior distribution of $(f_1, ..., f_m)$ is the Dirichlet distribution.

Several authors investigated nonparametric Bayesian estimation of $F(t)$ or $S(t) = 1 - F(t)$ using the Dirichlet process prior or its variant for interval-censored failure time data. Among them, Doss (1994) considered prior distributions on $F(t)$ that are a mixture of Dirichlets and give most of their mass to small neighborhoods of some parametric family. Calle and Gómez (2001) studied the same problem and employed the Dirichlet process prior. Their approach is a generalization of that given in Susarla and van Ryzin (1976), who derived nonparametric Bayesian estimators of $S(t)$ for right-censored data that include the Kaplan-Meier estimator as a special case. More recently, using the Dirichlet process prior, Zhou (2004) gave another nonparametric Bayesian estimator of $S(t)$ that has an explicit form. In contrast, the estimators given by both Doss (1994) and Calle and Gómez (2001) have no closed forms.

As mentioned above, the beta process prior is another prior commonly used in the analysis of failure time data. In fact, as pointed out in Ibrahim et al. (2001), the prior that is more convenient and often sufficient in practice is the discrete beta prior rather than the continuous beta prior. Consider a discrete survival variable T taking only the finite number of values denoted by $s_1 < ... < s_m$ as above. Let p_j denote the hazard of T at time s_j as defined in Section 1.4.2, $j = 1, ..., m$. The discrete beta prior assumes that the p_j's are independent and p_j follows the beta distribution $\mathcal{B}(c_j\alpha_j, c_j(1-\alpha_j))$, where the c_j's and α_j's are hyperparameters that are usually assumed to be known. In other words, under the discrete beta prior, the joint prior density function has the form

$$\pi(p_1, ..., p_m) = C \prod_{j=1}^{m} p_j^{c_j\alpha_j - 1} (1 - p_j)^{c_j(1-\alpha_j)-1}, \tag{10.10}$$

where C is the standardizing constant.

A useful feature of the discrete beta prior is that if exact survival data are available, the posterior distribution is the same as the prior distribution, but with different parameters. Specifically, let $\boldsymbol{O}_c$ denote the exact survival data and d_j and r_j the numbers of subjects for whom $T = s_j$ and who are at risk right before time s_j, respectively, $j = 1, ..., m$. Then we have

$$\pi(p_1, ..., p_m \,|\, \boldsymbol{O}_c) = C \prod_{j=1}^{m} p_j^{c_j\alpha_j + d_j - 1} (1 - p_j)^{c_j(1-\alpha_j)+r_j-d_j-1}.$$

Of course, for interval-censored failure time data, the d_j's and r_j's are unknown. Sinha (1997) studied the use of the discrete beta prior for Bayesian analysis of interval-censored data and discussed the selection of the hyperparameters c_j's and α_j's.

10.4.2 Semiparametric Bayesian Approaches

Consider regression analysis of the observed data $\boldsymbol{O}$ and let $\boldsymbol{\theta}$ represent regression parameters $\boldsymbol{\beta}$ and the baseline survival function S_0 together. In this case, a simple and easy choice for prior distribution is to let S_0 follow the Dirichlet process prior and to assume that $\boldsymbol{\beta}$ is independent of S_0 and has a multivariate normal distribution as its prior.

For discrete interval-censored failure time data, one can apply the same strategy as that with continuous interval-censored data. For example, assume that as above, the survival variable T of interest takes only finite values s_j's and for a subject with covariates $\boldsymbol{Z}$, the hazard at time s_j has the form

$$p_j(\boldsymbol{Z}) = p_j \, e^{\boldsymbol{Z}' \boldsymbol{\beta}} \, ,$$

where the p_j's denote the baseline hazards. That is, we have a piecewise constant hazard model. Then we can impose the prior given in (10.10) on the p_j's and an independent multivariate normal distribution on the regression parameters $\boldsymbol{\beta}$. Instead of the discrete beta prior, alternatively, one can apply the discrete version of the Gamma process prior (Ibrahim et al., 2001; Kalbfleisch, 1978), a commonly used nonparametric prior process for the PH model. Sinha et al. (1999) discussed this for situations with one covariate. Specifically, they assume that all the p_j's and β are independent and

$$p_j \sim \Gamma(c_j, \alpha_j) \ , \ \beta \sim N(\beta_0, \sigma_0^2) \, ,$$

or β depends on time and

$$\beta_{j+1} \,|\, (\beta_1, ..., \beta_j) \sim N(\beta_j, \sigma_k^2) \, ,$$

where β_j denotes the value of the regression parameter at time s_j. In this model, as before, the c_j's, α_j's, β_0 and σ_j's are hyperparameters and assumed to be known. Of course, one can use other models in Bayesian analysis (Hanson and Johnson, 2004).

10.4.3 Some Remarks

It is apparent that if there exists prior information about the survival variable of interest, Bayesian approaches provide natural tools for the analysis because one can formally incorporate the information into the analysis through $\pi(\boldsymbol{\theta})$ (Dunson et al., 2004). An example of prior information is given by historical data that may exist for the disease or survival variable under study. In addition, in some situations, one may simply prefer to apply Bayesian approaches as analysis tools assuming that an appropriate prior and a sampling technique are available. For example, this could be the case in a situation where one has to reply on asymptotic results of a frequentist approach, but the sample size is small.

As discussed above, a Bayesian analysis involves choosing a prior as well as developing a sampling procedure to sample from the posterior distribution. This section discussed the first step, but not the second step. In general, the development of a sampling algorithm can be more difficult and again we refer readers to the book by Ibrahim et al. (2001) for discussion on this topic. Also parametric Bayesian approaches were not considered, but references include Banerjee and Carlin (2004) and Gómez et al. (2004). In addition to interval-censored survival times, one can also apply Bayesian approaches to analyze bivariate interval-censored data (Groenewald and Mokgatlhe (2004), or data with interval-censored covariates. For the latter, Calle and Gómez (2005) considered the model (10.8) with the T_i^*'s observed exactly.

10.5 Analysis with Informative Interval Censoring

The statistical methods discussed so far assume that the observation process that generates interval censoring is independent of the variable of interest. That is, we have independent interval censoring described in Section 1.3.5. In general, the contribution of an interval-censored observation given by (1.1) to the likelihood has the form

$$L_S^*(L=l, R=r; T) \,=\, P(l \,<\, T \,\leq\, r \,|\, L=l, R=r)\, dG(l,r)\,, \qquad (10.11)$$

where dG denotes the joint probability or density function of (L, R). Under independent interval censoring, this can be replaced by

$$L_S(L=l, R=r; T) \,=\, P(l \,<\, T \,\leq\, r\,)\,, \qquad (10.12)$$

which has been used throughout the book and gives a much simplified likelihood function compared with that obtained from the term (10.11). That is, with independent censoring, one can ignore the censoring mechanism or the observation process.

The observation process that controls interval censoring can be generated in many different ways (Babineau, 2005). A natural question is under what conditions we have or what types of observation processes give independent interval censoring. More generally, under what conditions or for what types of observation processes can one use the term (10.12), instead of the term (10.11), in the construction of a likelihood function? Furthermore, suppose that such conditions or processes exist. Then in practice, one needs to know how to test if (10.12) is appropriate and how inferences can be made for situations where the term (10.12) is not valid.

Before we further discuss independent interval censoring and the use of the terms (10.11) and (10.12), it is helpful to briefly review right-censored failure time data regarding the same issues. In this case, as discussed in Section 1.1.3, an independent right censoring mechanism that is similar to the independent interval censoring mechanism is commonly assumed. As with the latter, the use of the former greatly simplifies the likelihood construction for the analysis of right-censored data. It is well-known that independent right censoring is not always appropriate, but in general it cannot be tested (Tsiatis, 1975) unless more information about the censoring process is available or one imposes some modeling assumptions on the process. Among others, Lagakos and Williams (1978) and Williams and Lagakos (1977) investigated conditions under which one can and cannot, respectively, ignore the right censoring mechanism and apply a simplified likelihood contribution that is similar to (10.12) for inferences. In particular, Williams and Lagakos (1977) derived a constant-sum condition that allows the use of the simplified likelihood contribution. Kalbfleisch and MacKay (1979) also studied the constant-sum condition. For the analysis of right-censored data when the censoring mechanism cannot be ignored or with dependent right censoring, some recent references include DiRienzo (2003), Huang and Wolfe (2002), Lin et al. (1996), Robins and Finkelstein (2000), Rotnitzky and Robins (1995), and Scharfstein and Robins (2002).

In Section 10.5.1, we first briefly discuss conditions under which one can apply the term (10.12) to the likelihood construction for inferences and related issues. We refer the type of interval censoring for which (10.12) is valid as *noninformative interval censoring*. In contrast, the interval censoring for which the term (10.12) is invalid is referred to as *informative interval censoring*. In other words, one can classify interval-censored data into three types according to three interval censoring mechanisms: data with independent, noninformative, and informative interval censorings, respectively. It is obvious that independent interval censoring is noninformative interval censoring. Sections 10.5.2 describes some ideas that can be applied to the analysis of failure time data with informative interval censoring and is followed by some general remarks in Section 10.5.3.

10.5.1 Noninformative Interval Censoring

One wonders if there exists an interval censoring that is noninformative, but not independent interval censoring. Betensky (2000) first investigated this for current status data and provided a positive answer. Following Betensky (2000), Oller et al. (2004) studied the same problem, but for general case II interval censored data. In both Betensky (2000) and Oller et al. (2004), a constant-sum condition, similar to that given in Williams and Lagakos (1977), is described that characterizes an observation process that gives noninformative interval censoring.

To describe the constant-sum condition for case II interval-censored data, consider an interval-censored observation given in (1.1) and let $F(t)$ denote the cumulative distribution function of T. Then Oller et al. (2004) define the constant-sum condition as

$$\int_{(\,(l,r]:t\in(l,r]\,)} \frac{P(L \in dl\,,\, R \in dr\,,\, T \in (L,R])}{P(T \in (l,r])} = 1$$

for any t such that $dF(t) \neq 0$. Furthermore, they show that if an observation process satisfies the constant-sum condition, the resulting interval censoring is noninformative. For such a process, called the original process, there exists an observation process that yields independent interval censoring and gives the same marginal distributions of T and (L, R) as those arising from the original observation process. In other words, the constant-sum condition is equivalent to the existence of an observation process that gives both an independent interval censoring and the same probability distribution for the observed data as the underlying true process (Lawless, 2004).

More importantly, one wants to know if it is possible to test if interval censoring is noninformative. As remarked before, unfortunately, this is not possible in general. For example, if the observed data are given in the format (1.1) without additional information about the observation process, one cannot test if the observation process and the underlying survival process are independent. This is similar to the case of right-censored data. For the case of current status data, Betensky (2000) gives an example of a class of observation processes that includes constant-sum observation processes as a subclass, but within which, the constant-sum condition cannot be tested.

As with right-censored data, one way to test for independent or noninformative interval censoring is to obtain more information about and/or impose some model assumptions on the underlying observation process. Betensky and Finkelstein (2002) discussed this for the situation where there exists a sequence of prespecified observation times for all subjects. Using the conditional model of failure given observation status at each observation time, they developed a procedure for testing a negative correlation between observation process and failure process.

10.5.2 Informative Interval Censoring

This subsection considers the analysis of interval-censored failure time data when interval censoring is informative. That is, when one cannot apply the term (10.12) for the construction of a likelihood function. Instead, we have to deal with the term (10.11) or the joint probability function or joint density function, $dG(l, r)$. For this, several ideas can be applied. One is to assume that there exists another variable that can be observed and conditional on which $dG(l, r)$ does not involve the parameters of interest. Of course, here we need to assume that the distribution of the variable that is conditioned on does not involve the parameters of interest. Among others, van der Laan and Hubbard (1997) and van der Laan and Robins (1998) considered this approach for the situation where a surrogate marker is available and serves as the conditional variable. This approach is essentially similar to those given in other sections of the book in that they avoid dealing directly with the interval censoring mechanism. Another idea is to conduct Bayesian analysis. For example, Dunson and Dinse (2002) developed some Bayesian models for multivariate current status data in the presence of informative censoring.

In the following, we focus on directly modeling the observation process as well as the survival process of interest for the analysis of informatively interval-censored data. For this, two general approaches are discussed below. One, first discussed in Finkelstein et al. (2002), bases inferences on a full likelihood function and the other, proposed in Zhang, Sun, and Sun (2005), applies estimating equations. For the former, we focus on nonparametric estimation of a distribution function, and for the latter, the focus is on regression analysis of current status data using the additive hazards model (1.7).

10.5.2.1 A Full Likelihood Approach

Consider a survival study that consists of n independent subjects in which each subject is observed at a subset of a sequence of prespecified time points $t_1 < ... < t_m$. That is, we have discrete interval-censored data and for the survival variable T of interest, only the probabilities $\{\, g_j = P(t_{j-1} < T \leq t_j) \,\}$ can be estimated. We assume that in the study, as in Sections 6.3 and 6.4, subjects continue to be observed even after the failure has occurred and observation times are recorded. For subject i, let l_i denote the last observation time at which the failure has not occurred and r_i the first observation time at which the failure has occurred. That is, $(l_i, r_i]$ denotes the interval to which the survival time T_i from the subject is observed to belong. For each (i, j), define $\delta_{ij} = 1$ if subject i is observed at time t_j and 0 otherwise and $\alpha_{ij} = 1$ if $(l_i, r_i]$ contains t_j and 0 otherwise, $i = 1, ..., n$, $j = 1, ..., m$. Then the full likelihood function is proportional to

$$L_S = \prod_{i=1}^{n} P(T_i \in (l_i, r_i] \,|\, \delta_{i1}, ..., \delta_{im}) \, d\,G_i(\delta_{i1}, ..., \delta_{im}) \tag{10.13}$$

in terms of the conditional distribution of the T_i's, or

$$L_S = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} \, g_j \, d\,G_i(\delta_{i1}, ..., \delta_{im} \,|\, T_i \in (t_{j-1}, t_j]) \tag{10.14}$$

in terms of the marginal distribution of the T_i's, where $d\,G_i(\delta_{i1}, ..., \delta_{im})$ and $d\,G_i(\delta_{i1}, ..., \delta_{im} \,|\, T_i \in (t_{j-1}, t_j])$ denote the marginal and conditional probability or density functions of the δ_{ij}'s, respectively.

Suppose that one is interested in estimation of the survival function of T or in particular, the g_j's. In this case, it is apparent that the likelihood given in (10.14) is more convenient than that given in (10.13) and one can maximize it with respect to all parameters. To derive an efficient estimate, we need some assumptions about $d\,G(\delta_{i1}, ..., \delta_{im} \,|\, T_i \in (t_{j-1}, t_j])$. For this, Finkelstein et al. (2002) assume that the distributions of the δ_{ij}'s are the same for all subjects and given T_i, the δ_{ij}'s are independent. Then the likelihood function given in (10.14) can be rewritten as

$$L_S = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}\, g_j \left\{ \prod_{k=1}^{j-1} \left[p_{1k}^{\delta_{ik}} (1 - p_{1k})^{1-\delta_{ik}} \right] \prod_{k=j}^{m} \left[p_{2k}^{\delta_{ik}} (1 - p_{2k})^{1-\delta_{ik}} \right] \right\} , \tag{10.15}$$

where $p_{1j} = P(\delta_{ij} = 1 \,|\, T > t_j)$ and $p_{2k} = P(\delta_{ij} = 1 \,|\, T \leq t_j)$, $j = 1, ..., m$. To maximize L_S in (10.15), Finkelstein et al. (2002) present an EM algorithm.

Suppose that information about some covariates is available and we are interested in estimating the effects of these covariates. In this case, one can derive a likelihood function similar to that given in (10.15) after specifying some regression models for the dependence of the g_j's, p_{1j}'s, and p_{2j}'s on the covariates, respectively.

We remark that a major assumption in the construction of the likelihood functions given in (10.13) to (10.15) is that the follow-up on study subjects continues until the end of study or subjects may drop out of the study, but not due to the occurrence of failure. Without these assumptions concerning follow-up, the method described would not apply. Another key assumption about the likelihood given in (10.15) is the conditional independence of the δ_{ij}'s given the survival variable. An alternative to this is to assume that the observation probability at each time point depends on the observation probability or status at the previous time point, or on how close the time point is to the survival time.

10.5.2.2 An Estimating Equation Approach

In this subsection, we consider regression analysis of current status data as in Chapter 5. Specifically, using the notation defined before, suppose that the data arising from n independent subjects have the form

$$\{ \, (C_i, \delta_i = I(T_i \leq C_i), \boldsymbol{Z}_i) \, ; \, i = 1, ..., n \, \} \, .$$

Unlike Chapter 5, we assume that the T_i's and C_i's may be correlated. Suppose that the goal is to make inferences about covariate effects.

We assume that the relationship between T_i and C_i can be characterized by a possibly time-dependent random effect $b_i(t)$, and given $b_i(t)$, T_i and C_i are independent. Furthermore, we assume that given $\boldsymbol{Z}_i$ and $b_i(t)$, the conditional hazard functions of T_i and C_i are given by, respectively,

$$\lambda(t; \boldsymbol{Z}_i, b_i(s), s \leq t) \, = \, \lambda_0(t) \, + \, \boldsymbol{Z}_i' \boldsymbol{\beta} \, + \, b_i(t) \tag{10.16}$$

and

$$\lambda_i^c(t; \boldsymbol{Z}_i, b_i(s), s \leq t) \, = \, \lambda_c(t) \, \exp(\boldsymbol{Z}_i' \boldsymbol{\gamma} \, + \, b_i(t)) \, . \tag{10.17}$$

This is the same notation used for models (1.4) and (5.9). It is apparent that if $b_i(t) = 0$ for all i, the problem considered here reduces to that discussed in Section 5.4. Also under the assumptions above, T_i and C_i are independent if $b_i(t) = 0$ or more generally if it has zero variability. This can be used to test for independence in interval censoring. As remarked before, one interesting and useful feature of the model (10.16) is that under it, the marginal hazard of the T_i's also follows the additive hazards model (1.7) (Lin, Oakes, and Ying, 1998; Zhang, Sun, and Sun, 2005).

The models (10.16) and (10.17) are special cases of those proposed in Zhang, Sun, and Sun (2005), who considered the same problem, but with time-dependent covariates. They show that under models (10.16) and (10.17), the estimation approach described in Section 5.4 applies here. As discussed in Section 5.4, this estimation approach is easily implemented, but could be less efficient than a full likelihood-based approach. Also, for this estimating equation approach, the types of correlations between T_i and C_i that can be described by models (10.16) and (10.17) can be limited. A simple generalization, which addresses this difficulty, replaces model (10.16) by

$$\lambda(t; \boldsymbol{Z}_i, b_i(s), s \leq t) \;=\; \lambda_0(t) \,+\, \boldsymbol{Z}_i' \boldsymbol{\beta} \,+\, \alpha\, b_i(t)\,,$$

where α is an unknown parameter that represents the direction of the correlation between T_i and C_i. For this approach, of course, one also needs to address model-checking issues.

10.5.3 Some Remarks

A field that has many similarities with interval-censored data is missing data, in which the mechanism that causes missingness also plays an important role in their analysis (Little and Rubin, 1987). In terms of missingness, interval-censored data can be seen as special cases of missing data. But for missing data, the value of the variable of interest is completely unknown, while interval-censored data do provide ranges about the missing values. As with interval-censored data, missing data can also be classified into three categories according to three commonly used missing mechanisms, missing completely at random, missing at random, and nonignorable. They are similar, respectively, to independent, noninformative, and informative interval censorings. Another concept that is often used in the literature and is similar to both missing at random and noninformative interval censoring is coarsening at random (Gill et al., 1997; Heitjan and Rubin, 1991; van der Laan and Robins, 1998). It was introduced by Heitjan and Rubin (1991) for general incomplete data to represent coarsening mechanisms that can be ignored in making inferences about the response variable of interest.

For the analysis of failure time data with informative interval censoring, in addition to the likelihood and estimating equation approaches described in the previous subsection, one can also apply other ideas discussed in the previous chapters. In general, one needs to develop realistic models that are appropriate for the survival process of interest and the observation process both individually and jointly. Of course, here it is assumed that necessary information is available to ensure that the models are identifiable.

As commented above, it is in general impossible to test if an interval censoring process is noninformative unless additional information about the process is available. The same is true for the analysis of informatively interval-censored data in that one may know some reasonable models for the underlying observation process, but no additional information exists and thus one cannot identify the specific model for the data. In this case, one may conduct a sensitivity analysis, which is commonly used in general data analysis when there exists some uncertainty that cannot be resolved with the available information.

10.6 Computational Aspects and Additional Topics

Computational aspects of the various inference approaches described in this book are definitely an important component of their applications but are beyond the scope of the book. Currently, there is no statistical software that provides an extensive coverage for the analysis of interval-censored failure time data. But one can find some functions in Splus and R and some procedures in SAS that apply to interval-censored data. Also, there exist some personal packages written for the analysis of interval-censored data in the literature. For example, one can use the function *kaplanMeier* in Splus or install the R package *Icens* for nonparametric estimation based on interval-censored data. Similar results can be obtained using the SAS procedure LIFETEST. Kooperberg and Stone (1992) provide Splus software for smooth estimation of a survival function based on interval-censored data using splines, and Fay (1999b) gives some Splus functions for nonparametric estimation and treatment comparison. For examples given in this book, most programming was done using Splus or R. References that provide lengthy discussion about computational aspects include Lindsey and Ryan (1998) and Gómez et al. (2004).

In addition to the topics discussed in the previous sections, several other topics concerning interval-censored data could occur in practice although there exists little research on them in the literature. One is the analysis of stratified interval-censored failure time data. By stratified interval-censored data, we mean that the data are a combination of several data sets arising from different sources or backgrounds but are generated or collected for the same purpose. For example, in a cancer study that aims to evaluate the effect of a new drug, the study may consist of both male and female patients. For the analysis, it is apparent that a simple way is to define a gender covariate to take into account the gender effect. In many situations, however, this may not be good enough. A better approach may be to apply some stratified regression models such as the stratified PH or additive hazards model, the model (1.4) or (1.7) with different baseline hazard functions, for male and female patients.

Competing risk problems often occur and have been extensively discussed in the context of right-censored failure time data (Kalbfleisch and Prentice, 2002). For the same reasons, one could face these problems when only interval-censored data are available. The analysis of interval-censored data that involve covariates with measurement errors is another topic that one may need to deal with. Also, we could have interval-censored data that arise from case-cohort studies, for which one may want to develop analysis methods that take into account the special features of the data or study.

This page intentionally blank

Appendix

Some Sets of Data

The following sets of data are used for examples and discussion at various places of the book.

Data set I, given in Table A.1, arises from ACTG 181 and is discussed in Section 1.2.4 and analyzed in Sections 7.2.3 and 7.4.3. The variables of interest are times to CMV shedding in blood and urine, respectively, and the available information includes observed intervals $(L_U, R_U]$ and $(L_U, R_U]$ for both the blood and urine shedding times. The variable CD4.ind indicates if the patient's baseline CD4 cell count is below 75 (cells/μl) (by 1).

Data set II, given in Table A.2, arises from a 16-center prospective study in the 1980s on people with hemophilia for the purpose of investigating the risk of HIV-1 infection on these people. It is analyzed in Sections 3.4.4 and 6.2.3. The table includes observed intervals within which the HIV-1 infection occurred for 368 patients who received no and low-dose factor VIII concentrate.

Data set III, given in Table A.3, arises from ACTG 359 and is analyzed in Sections 6.3.2 and 6.4.2. The table presents the observed information on 271 AIDS patients whose numbers of RNA copies were measured at least once during the 12-month period. In the table, for each patient and at each of 8 prescheduled time points, 1 means that an observation is available and the RNA number has already dropped below 500 at or before the time point, 0 means that an observation is available, but the RNA number has not yet dropped below 500 up to the time point, and a dot means no observation at the time point.

Data set IV, given in Table A.4, arises from the National Cooperative Gallstone Study, a 10-year, multicenter, double-blinded, placebo-controlled clinical trial of the use of the natural bile acid chenodeoxycholic acid (cheno) for the dissolution of cholesterol gallstones. It is analyzed in Sections 9.2.3 and 9.3.1. The table includes the successive visit times in study weeks and the associated counts of episodes of nausea for the 111 patients in the high-dose cheno and placebo groups during the first year of the study.

Data set V, given in Table A.5, arises from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group and is analyzed in Section 9.4.3. In the table, dot means no visit and the number represents the number of bladder tumors that occurred between the previous and current visits. The second column gives the size of the largest initial tumor, and the number of initial tumors (at month 0) is given in column 3.

Table A.1. Data set I — Observed intervals in weeks for blood and urine shedding times along with the baseline CD4 status from ACTG 181

Patient	L_B	R_B	L_U	R_U	$CD4.ind$	Patient	L_B	R_B	L_U	R_U	$CD4.ind$
1	11	-	11	-	1	45	6	10	0	2	1
2	11	-	11	-	1	46	2	-	2	-	1
3	11	-	11	-	0	47	13	-	13	-	0
4	11	-	8	10	0	48	15	-	0	3	0
5	7	-	6	8	1	49	8	-	0	1	0
6	11	-	12	-	0	50	16	-	6	9	0
7	8	12	8	10	1	51	5	-	0	1	1
8	10	-	10	-	0	52	2	-	0	1	1
9	6	-	6	-	1	53	13	-	0	1	0
10	2	9	9	11	0	54	13	-	13	-	1
11	11	-	11	-	1	55	5	-	0	1	1
12	5	-	0	1	1	56	5	-	9	-	1
13	16	-	16	-	1	57	17	-	0	1	1
14	0	1	0	1	1	58	8	-	1	3	1
15	18	-	1	2	0	59	16	-	3	5	1
16	16	-	0	1	1	60	0	1	0	1	1
17	16	-	0	1	0	61	2	6	0	1	1
18	10	-	10	-	1	62	14	-	13	-	0
19	10	-	0	1	1	63	16	-	17	-	0
20	19	-	1	3	0	64	13	-	13	-	0
21	2	14	9	11	1	65	1	-	1	-	1
22	0	3	1	3	1	66	14	-	0	3	1
23	18	-	18	-	0	67	12	-	6	9	0
24	11	-	9	-	0	68	0	1	0	1	1
25	15	-	1	5	0	69	13	-	0	1	1
26	5	-	1	2	1	70	16	-	3	4	0
27	11	14	12	14	0	71	15	-	0	1	0
28	13	-	0	1	1	72	16	-	5	7	1
29	14	-	16	-	0	73	8	-	0	1	1
30	0	1	1	2	1	74	1	2	1	2	1
31	16	-	17	-	0	75	1	4	0	1	1
32	16	-	1	2	1	76	15	-	1	2	1
33	16	-	17	-	0	77	5	9	0	1	1
34	5	9	4	6	1	78	16	-	16	-	0
35	1	-	1	-	0	79	1	2	0	1	1
36	9	-	4	-	0	80	15	-	1	3	1
37	12	-	11	-	0	81	14	-	0	1	0
38	2	-	1	4	1	82	8	-	1	4	0
39	10	-	11	-	1	83	15	-	2	4	0
40	8	-	11	-	0	84	13	-	13	-	0
41	12	-	12	-	1	85	4	-	0	1	1
42	12	-	12	-	0	86	13	-	9	13	1
43	5	-	11	-	0	87	4	-	0	1	1
44	2	-	3	-	1	88	5	-	4	6	1

Data set I (Continued)

Patient	L_B	R_B	L_U	R_U	$CD4ind$	Patient	L_B	R_B	L_U	R_U	$CD4ind$
89	12	-	12	-	1	133	13	-	13	-	0
90	13	-	13	-	0	134	1	3	0	1	1
91	1	-	13	-	0	135	11	-	5	7	0
92	8	-	9	-	1	136	11	-	4	5	1
93	9	-	11	-	0	137	2	-	1	3	1
94	1	-	1	-	1	138	8	-	0	1	1
95	13	-	13	-	0	139	13	-	13	-	0
96	2	-	4	-	1	140	11	-	0	1	0
97	13	-	1	3	1	141	13	-	13	-	0
98	3	-	5	-	1	142	16	-	0	1	1
99	10	-	10	-	0	143	11	-	4	6	1
100	9	-	9	-	1	144	1	-	0	1	1
101	6	-	6	-	1	145	8	-	15	-	0
102	1	-	10	-	1	146	8	14	9	10	1
103	1	-	1	-	1	147	13	-	12	15	0
104	11	15	10	14	0	148	8	-	16	-	1
105	5	9	0	1	0	149	8	-	7	10	0
106	19	-	0	1	1	150	2	11	11	13	0
107	16	-	1	2	0	151	11	-	11	-	0
108	20	-	1	3	1	152	5	-	10	-	0
109	17	-	17	-	0	153	6	-	6	-	1
110	1	4	0	1	0	154	14	-	14	-	0
111	15	-	0	1	0	155	0	1	0	1	1
112	19	-	19	-	0	156	13	-	0	1	0
113	19	-	6	8	0	157	14	-	14	-	1
114	14	-	6	9	1	158	13	-	13	-	0
115	19	-	5	7	0	159	13	-	14	-	0
116	19	-	1	2	1	160	13	-	13	-	1
117	17	-	17	-	0	161	14	-	6	7	1
118	8	-	3	6	1	162	14	-	14	-	0
119	9	-	15	-	0	163	12	-	9	11	0
120	19	-	1	3	0	164	12	-	13	-	1
121	18	-	18	-	0	165	7	-	9	-	1
122	1	6	1	6	1	166	14	-	1	2	1
123	19	-	19	-	0	167	14	-	14	-	0
124	1	3	8	11	0	168	13	-	3	5	1
125	11	15	1	4	1	169	13	-	13	-	1
126	0	1	0	1	1	170	14	-	12	-	0
127	17	-	17	-	0	171	11	-	13	-	0
128	11	-	2	6	1	172	12	-	13	-	1
129	11	14	2	6	1	173	8	-	8	-	1
130	19	-	0	1	0	174	12	-	8	9	0
131	1	-	0	1	1	175	4	-	4	-	0
132	1	4	0	4	1	176	11	-	2	4	1

Data set I (Continued)

Patient	L_B	R_B	L_U	R_U	$CD4ind$	Patient	L_B	R_B	L_U	R_U	$CD4ind$
177	11	-	1	3	0	191	1	-	1	-	1
178	14	-	12	14	1	192	6	-	6	-	1
179	19	-	19	-	0	193	10	-	0	1	0
180	19	-	1	3	1	194	1	-	0	1	1
181	16	20	0	1	0	195	8	-	1	2	1
182	1	-	1	-	1	196	3	-	1	2	1
183	10	-	8	10	1	197	13	-	13	-	1
184	6	-	0	1	0	198	14	-	1	2	1
185	13	-	1	2	0	199	13	-	13	-	0
186	11	-	11	-	0	200	11	-	11	-	0
187	11	-	6	8	0	201	7	-	7	-	1
188	2	-	0	1	1	202	14	-	0	1	1
189	10	-	1	3	1	203	14	-	14	-	0
190	11	-	0	1	1	204	4	-	4	-	0

Table A.2. Data set II — Observed intervals $(L_i, R_i]$ in quarters for HIV-1 infection times of 368 people with hemophilia in no and low-dose groups from a 16-center prospective study

Patient	L_i	R_i	Patient	L_i	R_i	Patient	L_i	R_i
			No factor VIII concentrate group					
1	55	-	80	29	-	159	47	-
2	55	-	81	54	-	160	47	-
3	56	-	82	53	-	161	49	-
4	54	-	83	16	19	162	21	27
5	53	-	84	48	-	163	0	25
6	57	-	85	55	-	164	46	-
7	31	33	86	29	-	165	27	-
8	56	-	87	46	-	166	49	-
9	56	-	88	55	-	167	27	33
10	54	-	89	54	-	168	18	34
11	56	-	90	53	-	169	45	-
12	54	-	91	53	-	170	49	-
13	55	-	92	20	24	171	54	-
14	56	-	93	53	-	172	6	31
15	57	-	94	54	-	173	47	-
16	56	-	95	51	-	174	0	43
17	54	-	96	48	-	175	57	-
18	56	-	97	18	22	176	34	-
19	54	-	98	51	-	177	54	-
20	5	30	99	32	-	178	0	31

Data set II (Continued)

Patient	L_i	R_i	Patient	L_i	R_i	Patient	L_i	R_i
21	54	-	100	20	32	179	56	-
22	55	-	101	18	24	180	56	-
23	57	-	102	54	-	181	25	31
24	55	-	103	17	22	182	35	-
25	56	-	104	33	-	183	50	-
26	52	-	105	47	-	184	56	-
27	46	-	106	53	-	185	0	35
28	45	-	107	33	-	186	55	-
29	49	-	108	48	-	187	54	-
30	55	-	109	47	-	188	25	40
31	57	-	110	20	28	189	48	-
32	54	-	111	38	-	190	42	-
33	57	-	112	4	33	191	55	-
34	55	-	113	41	-	192	53	-
35	43	-	114	52	-	193	52	-
36	48	-	115	55	-	194	56	-
37	40	-	116	26	31	195	56	-
38	54	-	117	52	-	196	56	-
39	51	-	118	24	28	197	56	-
40	56	-	119	53	-	198	55	-
41	56	-	120	54	-	199	55	-
42	51	-	121	54	-	200	54	-
43	55	-	122	50	-	201	57	-
44	54	-	123	53	-	202	39	-
45	35	-	124	54	-	203	56	-
46	35	-	125	37	-	204	57	-
47	56	-	126	55	-	205	53	-
48	42	-	127	48	-	206	30	-
49	54	-	128	51	-	207	56	-
50	45	-	129	54	-	208	57	-
51	56	-	130	45	-	209	56	-
52	56	-	131	48	-	210	55	-
53	56	-	132	50	-	211	55	-
54	50	-	133	0	41	212	57	-
55	54	-	134	41	-	213	30	-
56	39	-	135	47	-	214	5	30
57	56	-	136	41	-	215	0	29
58	51	-	137	51	-	216	44	-
59	55	-	138	46	-	217	30	-
60	50	-	139	50	-	218	30	-
61	45	-	140	47	-	219	56	-
62	49	-	141	47	-	220	31	-
63	52	-	142	47	-	221	55	-
64	52	-	143	54	-	222	56	-

Data set II (Continued)

Patient	L_i	R_i	Patient	L_i	R_i	Patient	L_i	R_i
65	57	-	144	53	-	223	54	-
66	56	-	145	53	-	224	54	-
67	55	-	146	54	-	225	57	-
68	14	49	147	54	-	226	54	-
69	54	-	148	43	-	227	38	-
70	55	-	149	49	-	228	54	-
71	55	-	150	44	-	229	54	-
72	54	-	151	29	-	230	57	-
73	56	-	152	46	-	231	57	-
74	55	-	153	40	-	232	44	-
75	55	-	154	9	27	233	55	-
76	53	-	155	49	-	234	55	-
77	0	4	156	45	-	235	56	-
78	54	-	157	49	-	236	54	-
79	20	24	158	49	-			
Low-dose factor VIII concentrate group								
237	7	20	281	25	34	325	30	-
238	9	20	282	53	-	326	45	-
239	0	25	283	41	-	327	21	26
240	57	-	284	50	-	328	16	32
241	23	26	285	0	36	329	17	24
242	8	21	286	0	29	330	49	-
243	20	26	287	55	-	331	0	37
244	25	27	288	0	55	332	0	41
245	24	29	289	10	16	333	0	30
246	12	21	290	13	29	334	56	-
247	26	29	291	14	19	335	0	30
248	54	-	292	0	16	336	55	-
249	18	22	293	11	29	337	51	-
250	14	22	294	11	20	338	0	30
251	11	17	295	31	-	339	50	-
252	55	-	296	40	-	340	45	-
253	8	15	297	53	-	341	8	30
254	29	31	298	11	15	342	5	30
255	55	-	299	20	24	343	53	-
256	57	-	300	15	20	344	11	41
257	15	20	301	32	-	345	52	-
258	18	22	302	54	-	346	3	33
259	14	22	303	51	-	347	0	47
260	56	-	304	33	-	348	7	49
261	23	30	305	17	26	349	56	-
262	17	21	306	14	17	350	57	-
263	54	-	307	41	-	351	6	29
264	20	31	308	42	-	352	7	29
265	56	-	309	53	-	353	55	-

Data set II (Continued)

Patient	L_i	R_i	Patient	L_i	R_i	Patient	L_i	R_i
266	23	27	310	0	26	354	8	29
267	56	-	311	49	-	355	7	29
268	53	-	312	39	-	356	36	-
269	15	19	313	18	29	357	7	28
270	52	-	314	22	25	358	55	-
271	0	17	315	50	-	359	49	-
272	0	21	316	54	-	360	46	-
273	46	-	317	38	-	361	0	30
274	16	23	318	20	30	362	57	-
275	24	32	319	46	-	363	30	-
276	16	24	320	51	-	364	53	-
277	53	-	321	6	30	365	12	21
278	12	20	322	53	-	366	56	-
279	18	22	323	0	30	367	38	-
280	0	33	324	45	-	368	0	44

Table A.3. Data set III — Observation times in months and the status of RNA numbers for 271 AIDS patients from ACTG 359

Observation times								Observation times							
1	2	3	4	6	8	10	12	1	2	3	4	6	8	10	12
109 patients with initial RNA < 20000															
0	1	1	1	1	.	1	.	1	1	1	1	1	1	1	1
0	0	0	.	.	.	.	.	0	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	1	1	1	1	1	1	1	1
1	1	1	1	1	.	.	.	0	0	1	1	1	1	1	.
.	1	1	1	1	1	.	1	0	0	0	0	0	.	.	.
1	1	.	.	.	.	.	.	0	1	1	1	1	.	.	.
1	1	1	1	1	1	1	1	0	0	0	0	0	.	.	.
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	1	.	1	1	.	.	.	1	1	1	1	1	.	.	.
0	1	1	1	1	.	.	.	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1	.	1
1	.	.	.	.	.	.	.	1	1	.	1	1	1	1	1
1	1	1	1	1	1	.	.	.	.	.	1	1	.	.	.
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	.	1	1	1	.	.
0	0	0	0	0	.	.	.	1	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	1	1	1	1	.	1	1	1
0	0	1	1	1	1	1	1	.	1	1	1	1	1	1	1
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Data set III (Continued)

Observation times								Observation times							
1	2	3	4	6	8	10	12	1	2	3	4	6	8	10	12
0	0	0	.	0	.	.	.	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	0	0	0	0	0	.	.	.
0	0	0	0	0	0	.	0	1	1	1	1	1	1	.	1
.	0	.	0	0	.	.	0	1	1	1	1	1	1	.	.
0	0	0	0	0	0	0	.	0	0	0	0	0	.	.	.
1	1	1	1	1	1	1	1	0	.	0	0	0	0	.	.
0	.	0	0	0	.	.	.	1	1	1	1	1	.	.	.
1	1	1	1	1	1	1	1	0	0	0	0	0	.	.	.
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	.
0	.	.	1	1	.	.	.	0	0	0	0	0	0	0	0
1	1	1	1	.	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	1	1	1	1	1	.	.	.
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	1	.	1	1	1	1	1
0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1
0	1	1	1	1	1	1	1	0	0	1	1	1	.	.	.
0	0	0	0	0	.	.	.	1	1	1	1	1	1	1	1
0	0	0	.	1	.	.	.	0	0	0	0	1	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	.	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	.	0	0	.	.	.	.	.	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.
.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	.	1	.	.	.	.	1	1	1	1	1	1	1	1
1	1	1	1	.	1	.	.	0	0	0	.	0	0	0	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	.	.	.	.	.	.	.	.	0	0	0	.	.	.	.
1	1	1	1	1	.	.	.	1	1	1	1	.	1	1	1
0	0	0	0	0	0	0	0	0	0	1	1	.	1	1	1
1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
0	0	0	0	.	.	.	.	1	1	1	1	1	1	.	1
0	0	0	0	0	.	.	.	1	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	0	0	0	0	1	.	.	.
1	1	1	1	1	1	1	1								
162 patients with initial RNA $\geq$ 20000															
0	1	1	1	1	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	0	0	0	0	0	0	0	.	.	.
0	0	0	0	0	0	1	.	1	1	1	1	1	1	.	1
1	1	1	1	1	1	1	1	0	0	0	0	0	.	.	.
0	0	0	0	0	0	1	1	0	0	0	0	0	.	.	.
0	.	.	.	.	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	0	0	.	0	0	0	0	.	.	.

Data set III (Continued)

Observation times								Observation times							
1	2	3	4	6	8	10	12	1	2	3	4	6	8	10	12
.	.	0	0	.	.	.	.	0	0	0	0	0	.	.	.
0	.	.	.	.	.	.	.	0	0	0	0	1	.	.	.
1	1	1	1	1	1	1	.	0	0	0	0	0	.	.	.
1	1	1	1	1	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	.	0	0	0	.	.	.
0	1	1	1	1	1	1	1	0	0	0	0	0	1	1	.
0	0	0	0	0	.	.	.	0	.	0	.	.	.	.	.
0	0	0	0	0	.	.	.	0	0	0	1	1	1	1	.
0	0	0	0	0	.	.	.	0	1	.	1	1	.	1	1
0	0	0	0	0	.	.	.	0	0	0	.	0	.	0	.
0	0	1	1	1	1	1	1	0	0	1	.	1	1	1	1
1	1	.	1	1	1	1	1	0	.	1	1	.	1	1	1
0	0	1	1	1	1	1	1	0	.	0	0	.	0	.	.
0	0	.	0	.	.	.	.	0	0	0	0	0	.	.	.
0	0	.	.	.	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	1	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	1	1	.	1	1	1
0	0	0	0	0	.	.	.	.	0	0	0	0	.	.	.
0	1	1	1	.	1	1	1	0	1	1	1	1	1	1	.
0	0	0	.	0	.	.	.	1	1	.	1	1	1	1	1
1	1	1	1	1	1	1	1	0	.	0	0	0	0	0	0
0	0	0	0	0	0	.	.	0	0	0	0	0	0	.	1
0	0	0	1	.	1	1	1	0	0	0	0	.	0	0	.
0	0	0	0	0	.	.	.	0	0	0	0	.	.	.	.
0	1	1	1	1	1	1	1	0	0	0	0	.	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	1	1	1	1	1	1
0	0	0	0	0	.	.	.	0	0	0	0	0	1	.	.
1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1
1	1	1	1	1	.	.	.	0	.	.	.	.	.	.	.
0	0	0	0	0	.	.	.	0	0	0	1	1	1	1	.
0	0	0	0	0	0	0	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	0	.	0	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	1	1	1	.	1	0	0	0	0	.	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
.	1	1	1	1	1	1	.	1	1	1	1	1	1	1	1
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	.	0	0	.	.	.	.	0	0	0	0	.	.	.	.
.	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	.	0	0	0	0	0	0	0	0	0	0	0
.	.	0	0	.	.	.	.	1	1	1	1	1	1	1	1

Data set III (Continued)

Observation times								Observation times							
1	2	3	4	6	8	10	12	1	2	3	4	6	8	10	12
.	0	0	1	1	1	1	1	0	0	0	0	0	.	.	.
1	1	1	1	1	1	1	1	0	0	0	0	0	.	.	.
0	0	0	0	.	.	.	.	0	0	0	0	0	0	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	0	.	.
0	1	1	1	1	1	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	.	.	0	.	0	1	1	.	.	.
.	1	.	1	1	1	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	1	1	1	1	1	1	1	0	0	1	1	.	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	0	.	.
0	0	0	0	.	.	.	.	0	0	.	0	0	.	.	.
0	0	0	0	0	.	.	.	0	1	1	1	1	.	.	.
.	0	0	0	0	.	.	.	0	0	1	1	1	1	1	1
0	1	1	1	.	.	.	.	0	0	0	0	0	1	.	.
0	0	0	0	0	.	.	.	0	1	1	1	1	1	1	1
.	0	0	0	.	.	.	.	0	0	0	0	0	.	.	.
1	1	1	.	.	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	.	1	1	1	.	.	.
0	0	0	0	0	.	.	.	0	.	0	0	0	.	.	.
0	0	0	0	.	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	0	.	0	0	1	1	1	1	1	1
0	0	.	0	.	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	0	.	.	0	0	0	0	0	.	.	.
.	0	0	0	0	.	.	.	0	0	1	1	1	1	1	1
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.
0	0	0	0	1	1	1	1	0	0	.	0	0	.	.	.
0	0	0	0	0	.	.	.	1	1	1	1	1	1	1	1
1	1	1	1	1	.	.	.	0	0	0	0	0	0	0	0
0	0	0	0	0	.	.	.	0	0	0	0	0	.	.	.

Table A.4. Data set IV — Visit times in weeks and observed counts of episodes of nausea for 111 patients with floating gallstones in the National Cooperation Gallstone Study

Patient	Visit times and episodes of nausea																			
ID	t_1	N_1	t_2	N_2	t_3	N_3	t_4	N_4	t_5	N_5	t_6	N_6	t_7	N_7	t_8	N_8	t_9	N_9	t_{10}	N_{10}
	High-dose cheno group																			
1	4	0	8	0	13	0	26	0	38	0	51	0	69	0	.	.	.	.	.	.
2	4	0	9	3	13	0	26	0	39	0	51	0	68	0	.	.	.	.	.	.
3	4	0	8	0	12	0	24	0	38	0	51	0	69	0	.	.	.	.	.	.
4	4	0	8	0	12	0	26	0	38	0	51	0	69	0	.	.	.	.	.	.
5	4	0	8	0	13	0	26	0	38	0	52	0	70	0	.	.	.	.	.	.
6	4	0	8	0	12	0	25	0	39	0	51	0	68	0	.	.	.	.	.	.
7	4	0	9	0	14	0	26	0	39	0	52	0	70	0	.	.	.	.	.	.
8	4	0	9	0	14	0	28	0	39	0	53	0	69	0	.	.	.	.	.	.
9	4	0	9	1	14	0	27	1	38	1	54	4	71	0	.	.	.	.	.	.
10	4	0	9	0	13	0	17	0	22	0	26	0	38	0	43	0	62	0	.	.
11	3	0	8	0	13	0	26	0	40	4	53	2	68	0	.	.	.	.	.	.
12	4	0	8	0	13	1	27	0	39	0	52	0	70	0	.	.	.	.	.	.
13	4	20	10	2	14	2	17	10	28	0	41	0	54	6	71	0	.	.	.	.
14	5	1	9	0	13	0	26	0	38	0	52	0	69	0	.	.	.	.	.	.
15	5	0	9	0	15	0	27	0	39	0	51	0	70	0	.	.	.	.	.	.
16	4	0	9	0	13	0	26	0	38	0	52	0	69	0	.	.	.	.	.	.
17	4	0	8	0	12	0	27	0	39	0	51	0	68	0	.	.	.	.	.	.
18	4	0	8	0	12	0	26	0	37	0	48	0	70	0	.	.	.	.	.	.
19	4	0	9	0	14	0	28	0	38	0	52	0	71	0	.	.	.	.	.	.
20	9	0	22	0	31	0	38	0	55	0	68	0	.	.	.	.	.	.	.	.
21	5	0	10	0	13	0	25	0	50	2	81	0	.	.	.	.	.	.	.	.
22	4	0	9	0	12	0	25	0	39	0	50	0	59	0	81	0	.	.	.	.
23	5	0	8	0	13	0	25	0	40	0	60	0	.	.	.	.	.	.	.	.
24	4	0	9	0	13	0	26	0	38	0	51	0	69	0	.	.	.	.	.	.
25	4	0	9	0	13	0	26	0	38	0	52	99	84	0	.	.	.	.	.	.
26	4	0	9	1	13	0	26	0	39	0	53	5	68	2	.	.	.	.	.	.
27	3	0	8	0	13	1	25	0	40	0	51	0	61	0	.	.	.	.	.	.
28	4	0	8	0	13	0	24	0	38	0	52	0	68	0	.	.	.	.	.	.
29	3	0	9	0	12	5	26	0	38	0	50	0	60	0	.	.	.	.	.	.
30	4	0	10	0	15	1	28	0	41	0	55	3	72	0	.	.	.	.	.	.
31	3	0	8	0	13	0	26	0	39	0	52	0	93	0	.	.	.	.	.	.
32	3	1	9	3	13	0	26	0	38	0	52	0	70	0	.	.	.	.	.	.
33	4	0	10	0	16	0	29	0	41	0	54	6	72	0	.	.	.	.	.	.
34	3	0	7	0	12	0	25	0	38	0	51	0	71	0	.	.	.	.	.	.
35	4	0	9	0	13	0	26	0	39	0	51	0	69	0	.	.	.	.	.	.
36	5	0	9	2	13	0	26	0	39	0	51	0	68	0	.	.	.	.	.	.
37	6	0	12	6	16	0	28	0	41	0	63	0	.	.	.	.	.	.	.	.
38	4	0	9	0	13	0	25	0	38	0	51	0	70	0	.	.	.	.	.	.
39	4	0	8	0	12	0	26	0	40	0	53	0	71	0	.	.	.	.	.	.
40	4	0	8	0	12	10	26	0	39	0	52	0	71	5	.	.	.	.	.	.

Data set IV (Continued)

Patient	Visit times and episodes of nausea																			
ID	t_1	N_1	t_2	N_2	t_3	N_3	t_4	N_4	t_5	N_5	t_6	N_6	t_7	N_7	t_8	N_8	t_9	N_9	t_{10}	N_{10}
41	5	0	9	0	14	0	27	0	39	0	52	0	72	3	.	.	.	.	.	.
42	5	0	9	0	13	0	26	0	36	2	38	0	51	0	67	0	.	.	.	.
43	4	0	10	0	14	0	26	0	39	0	53	0	71	0	.	.	.	.	.	.
44	4	0	9	0	16	2	28	4	39	0	51	0	69	0	.	.	.	.	.	.
45	5	0	10	0	15	0	29	0	40	0	55	0	71	0	.	.	.	.	.	.
46	4	0	9	0	13	0	26	0	37	0	51	0	70	0	.	.	.	.	.	.
47	4	0	8	0	13	0	26	0	38	0	51	0	69	0	.	.	.	.	.	.
48	5	0	10	0	13	0	25	0	39	0	53	0	69	0	.	.	.	.	.	.
49	3	0	7	0	13	2	25	0	36	5	49	3	68	7	.	.	.	.	.	.
50	3	0	8	0	13	0	25	8	37	20	53	0	73	0	.	.	.	.	.	.
51	6	0	9	0	13	0	26	0	40	0	51	0	72	0	.	.	.	.	.	.
52	5	0	8	0	12	0	25	0	38	0	51	0	69	0	.	.	.	.	.	.
53	4	0	8	0	13	0	25	0	41	0	53	0	71	1	.	.	.	.	.	.
54	4	0	8	0	15	0	27	0	40	0	51	10	68	0	.	.	.	.	.	.
55	4	0	8	1	12	0	27	0	41	0	53	2	56	0	62	0	.	.	.	.
56	5	0	12	0	16	0	29	0	41	0	52	0	71	0	.	.	.	.	.	.
57	5	0	11	4	16	0	30	5	44	24	51	40	82	30	.	.	.	.	.	.
58	3	0	9	0	14	0	26	0	.	.	.	.	.	.	.	.	.	.	.	.
59	4	0	9	0	13	0	25	0	38	0	.	.	.	.	.	.	.	.	.	.
60	4	0	8	0	14	0	18	0	20	0	.	.	.	.	.	.	.	.	.	.
61	4	0	8	0	13	0	17	0	23	0	27	0	32	0	.	.	.	.	.	.
62	3	0	10	0	26	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
63	8	5	19	0	28	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
								Placebo group												
64	4	0	8	0	12	0	25	0	38	0	52	0	68	0	.	.	.	.	.	.
65	4	0	8	0	13	0	27	0	40	0	44	0	53	0	69	0	.	.	.	.
66	4	0	11	0	14	0	26	0	39	0	52	0	69	0	.	.	.	.	.	.
67	4	0	9	0	12	0	25	0	40	0	52	0	70	0	.	.	.	.	.	.
68	4	0	8	0	14	0	27	0	40	0	52	0	69	0	.	.	.	.	.	.
69	5	1	9	0	13	0	26	1	40	0	53	0	69	0	.	.	.	.	.	.
70	4	0	8	0	13	0	24	0	37	0	50	0	67	0	.	.	.	.	.	.
71	4	1	9	0	14	4	28	3	41	1	54	1	71	2	.	.	.	.	.	.
72	3	0	9	0	13	0	25	0	38	0	50	0	67	1	.	.	.	.	.	.
73	5	0	9	0	13	0	27	0	38	0	51	0	69	0	.	.	.	.	.	.
74	4	0	8	0	13	0	27	0	38	0	51	0	67	0	.	.	.	.	.	.
75	4	3	9	0	14	0	25	0	39	0	51	0	69	1	.	.	.	.	.	.
76	3	8	8	0	11	1	17	4	24	0	38	2	42	0	46	0	51	20	61	1
77	4	0	9	0	13	0	25	0	39	0	51	0	68	0	.	.	.	.	.	.
78	4	0	8	0	13	0	24	0	38	0	51	0	69	0	.	.	.	.	.	.
79	4	0	9	0	13	0	26	0	40	0	51	0	68	0	.	.	.	.	.	.
80	4	0	9	0	14	0	28	0	40	0	51	0	71	0	.	.	.	.	.	.
81	5	0	8	0	16	0	28	0	36	0	55	0	81	0	.	.	.	.	.	.
82	5	0	7	0	12	0	25	2	38	0	53	1	72	0	.	.	.	.	.	.

Data set IV (Continued)

Patient	Visit times and episodes of nausea																			
ID	t_1	N_1	t_2	N_2	t_3	N_3	t_4	N_4	t_5	N_5	t_6	N_6	t_7	N_7	t_8	N_8	t_9	N_9	t_{10}	N_{10}
									Placebo group											
83	5	0	10	0	15	0	29	0	41	0	55	0	69	0	.	.	.	.	.	.
84	4	0	9	0	13	0	25	0	35	0	56	0	74	0	.	.	.	.	.	.
85	4	0	9	0	13	0	28	0	39	0	59	0	70	0	.	.	.	.	.	.
86	4	0	9	3	12	0	24	0	37	0	51	0	68	0	.	.	.	.	.	.
87	4	0	8	60	13	0	24	0	40	1	55	0	74	4	.	.	.	.	.	.
88	3	0	8	1	14	0	26	0	38	0	53	0	70	0	.	.	.	.	.	.
89	5	0	9	0	13	0	27	0	40	0	54	0	73	0	.	.	.	.	.	.
90	3	0	8	0	11	0	25	0	37	0	51	0	68	0	.	.	.	.	.	.
91	3	1	7	4	11	0	24	0	38	0	54	9	.	.	.	.	.	.	.	.
92	3	5	8	0	13	0	25	0	38	0	52	0	68	0	.	.	.	.	.	.
93	4	0	9	0	13	0	26	3	39	0	52	0	70	0	.	.	.	.	.	.
94	4	0	9	0	14	0	26	0	39	0	52	0	68	0	.	.	.	.	.	.
95	4	6	9	0	18	1	28	0	39	0	54	0	74	10	.	.	.	.	.	.
96	5	0	9	0	15	0	27	0	39	0	53	0	69	0	.	.	.	.	.	.
97	4	0	9	0	13	2	25	0	38	0	50	0	68	0	.	.	.	.	.	.
98	3	3	7	0	12	0	25	6	38	0	52	0	69	0	.	.	.	.	.	.
99	4	0	7	0	12	0	25	0	38	1	53	0	69	0	.	.	.	.	.	.
100	4	0	8	0	13	0	26	0	39	0	51	0	70	0	.	.	.	.	.	.
101	4	0	8	0	13	0	26	0	40	0	52	0	78	10	.	.	.	.	.	.
102	4	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
103	4	0	8	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
104	5	0	9	0	13	0	17	0	21	0	28	1	39	1	.	.	.	.	.	.
105	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
106	6	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
107	3	25	8	30	14	20	.	.	.	.	.	.	.	.	.	.	.	.	.	.
108	4	0	9	0	13	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.
109	4	0	9	0	13	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	5	0	9	0	14	0	26	0	.	.	.	.	.	.	.	.	.	.	.	.
111	4	0	9	0	14	0	25	0	.	.	.	.	.	.	.	.	.	.	.	.

Table A.5. Data set V — Observed numbers of bladder tumors along with the numbers of initial tumors and the size of the largest initial tumor from a bladder cancer study

Patient	Size	Months																														
ID		0										10										20										30
Placebo group																																
1	3	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	1	2	0	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	1	1	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	1	5	.	.	0	.	.	.	.	.	0	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	1	4	0	.	.	0	.	1	0	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	1	1	.	.	0	.	.	.	.	.	.	0	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	1	1	.	0	.	.	.	.	.	.	.	0	.	2	.	.	.	3	.	0	.	.	.	.	.	.	.	.	.	.	.	.
8	1	1	.	.	0	.	.	.	.	.	.	.	.	.	.	0	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.
9	3	1	.	.	.	.	2	.	.	.	.	0	.	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.
10	3	1	.	.	0	.	.	.	0	.	.	6	.	.	.	.	3	.	.	.	0	.	.	.	0	.	.	.	.	.	.	.
11	1	1	0	.	8	.	.	.	.	0	.	0	.	.	0	.	.	8	.	.	0	.	.	.	8	.	.	.	.	.	.	.
12	1	3	.	.	1	.	.	0	.	.	1	0	.	.	0	.	0	.	0	.	.	0	8	.	0	.	.	.	.	.	.	.
13	3	3	.	.	0	.	.	0	.	.	0	.	0	.	.	0	.	.	.	0	.	.	.	.	0	.	.	.	.	.	.	.
14	3	2	.	.	0	.	.	.	8	.	.	7	.	.	0	.	.	5	.	.	.	.	.	.	.	7	.	.	.	.	.	.
15	1	1	.	.	1	.	.	0	.	.	0	.	.	0	.	.	1	.	0	.	.	0	.	0	.	.	3	.	.	.	.	.
16	1	8	8	.	.	0	.	.	0	.	.	.	0	.	.	0	.	.	0	.	.	0	.	.	.	.	0	0	.	.	.	.
17	4	1	.	4	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	.	.	.
18	2	1	.	.	0	.	.	0	.	.	.	.	.	.	0	.	.	0	.	.	.	.	.	0	0	.	.	0	.	.	.	.
19	2	1	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	0	.	.
20	4	1	.	.	.	.	0	.	.	0	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	0	.
21	2	1	.	.	0	.	.	0	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.
22	1	4	.	0	.	.	0	.	.	0	.	.	0	.	.	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	0	.
23	5	1	.	4	.	.	.	.	.	.	.	0	.	.	.	.	.	.	2	.	.	.	.	4	0	.	.	.	.	0	.	0
24	1	2	.	.	1	.	.	3	.	3	.	.	.	3	0	.	0	0	0	0	.	.	.	0	0	.	.	3	.	.	.	0
25	6	1	.	0	.	.	0	0	.	.	0	.	.	.	0	.	.	.	.	.	.	.	.	0	.	.	.	.	.	2	.	1
26	3	1	0	.	0	0	.	0	.	.	0	.	.	2	.	.	3	.	.	0	.	.	.	.	.	1	.	.	.	0	.	.
27	2	1	.	.	0	.	.	0	.	.	0	.	.	0	.	0	.	.	0	.	.	0	.	.	.	.	.	0	.	.	.	.
28	1	2	.	.	.	0	.	.	.	0	.	0	.	.	.	0	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	0
29	1	2	.	.	0	.	.	0	.	.	0	.	0	.	.	0	.	.	0	.	.	0	.	.	0	.	.	.	.	.	.	0
30	1	3	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	8	.
31	2	1	0	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	0	.	.	.	.	.	.	0	.	.	.	.	.	.
32	1	4	.	0	.	.	.	.	.	.	8	.	.	.	.	0	.	.	2	.	.	.	.	5	.	1	.	.	.	0	.	.
33	1	5	.	0	.	0	0	.	.	0	.	.	0	.	.	.	0	1	.	.	8	.	.	.	1	.	.	0	.	.	2	.
34	2	1	.	.	0	.	.	0	.	.	0	.	0	.	.	0	.	.	0	.	.	0	.	.	0	.	.	0	.	.	0	.
35	1	1	.	.	3	.	0	.	0	.	.	0	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	0	.	.	0
36	6	2	.	.	0	.	.	1	0	0	.	.	0	.	.	0	.	0	0	.	.	0	.	.	0	.	.	.	.	.	.	.
37	1	2	.	.	5	.	0	3	.	.	4	.	.	.	.	0	.	.	0	.	.	0	.	.	0	.	.	.	0	.	.	0
38	1	1	.	0	.	.	.	0	.	.	1	.	3	.	0	.	.	.	0	.	.	1	.	.	0	.	.	4	.	.	.	3
39	1	1	.	.	0	.	.	0	.	.	0	.	0	.	.	.	0	.	.	1	.	.	.	0	.	.	0	.	.	0	.	.
40	3	1	.	.	.	.	.	0	.	0	.	.	.	0	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	0
41	1	3	.	0	.	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.
42	7	1	.	.	0	.	.	.	.	.	.	0	.	.	0	0	.	.	1	.	.	.	0	.	.	0	.	.	0	.	.	0

Data set V (Continued)

Patient	Size	Months
ID		0 10 20 30
43	1	3 . . 7 . . . . 0 . . . 2 . . . . 0 . . . .
44	1	1 0 . . 0 . . 0 . . 0 . . 0 . . 0 . . 0 . . 0 . . . . . 0 . .
45	2	3 . 1 . . 0 . 0 . . 0 . . 0 . 3 . . . 0 . . . . 4 . . 0 . . 3
46	3	1 . 0 . . 3 . . 0 . . 0 . . 4 . . . 0 2 0 . . 0 . . . 5 . 0 .
47	3	2 . 1 . . 0 . . 3 . . . 6 2 . . . 2 . . . 1 . 0 0 . . 0 . . 0
		Thiotepa group
48	3	1 0 . . . .
49	1	1 0 . . . .
50	1	8 . . . . 8 . . . .
51	2	1 0 0 0 0 0 0 0 0 0 . . . .
52	1	1 . . . . . 0 . . . 0 . . . .
53	1	1 . . 0 . . . . 0 . . . .
54	6	2 . . 1 . 0 . . . 0 . 0 . . 0 . . . .
55	3	5 5 . 2 . 5 . 2 . . 2 . . 0 . . 0 0 . . . .
56	3	1 . . . . 0 . . . . . 0 . . . . . 2 0 . . . .
57	1	5 . . . . 0 . . . .
58	1	5 0 2 0 . . . . 0 . . . .
59	1	1 0 0 0 0 0 0 0 0 0 . . . 0 0 0 0 1 . 1 . 0 . . . .
60	1	1 . 0 . . 0 . . . . 0 . . . . 0 . . 0 . . . 0 . . . .
61	3	1 0 0 0 . . . . 0 . . . 0 . . . . . 0 . . 0 . . . 0 . . . . .
62	5	1 0 . . . . 0 . 0 . . . . 0 . . . . . 0 . . . . .
63	1	1 0 0 0 0 0 0 0 0 0 0 . 0 . 0 0 . . 0 0 . . 0 0 . 0 . . . . .
64	1	1 0 0 0 0 0 2 . . . 0 0 3 1 . . . . . . 0 . . 0 . . 0 . . . .
65	1	1 0 . 0 . . 1 . . . 0 . . 0 . . 0 . . . . 0 . . . . . 0 . . .
66	1	2 . 2 . . . . 0 .
67	3	8 . . 0 . . 0 . . 0 . . . . 0 . 0 0 . . 3 . . 0 .
68	1	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . 0 0 0 0 . 0 0 0 . 0 0 . . . .
69	1	6 0 0 0 1 . . 0 0 0 0 0 0 0 0 . 3 . 0 0 . 0 0 3 . 0 0 3 . 0 .
70	1	1 0 0 0 0 0 0 0 0 0 0 0 0 . 0 0 0 . 0 0 0 . 2 1 0 0 . 2 0 . .
71	1	3 . . . . 0 . . . . 0 . . 0 . . . . . . 3 . 2 . . 1 .
72	2	3 0 0 0 0 0 0 0 0 . 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . 0 0 0 0 0 0
73	1	1 0 0 0 0 . . 0 0 0 0 0 0 0 . 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 .
74	1	1 1 . 0 . . . 0 . . . . . 0 . . . 0 0 0 . 0 0 0 0 . 0 1 . 0 0
75	1	1 . . . . 0 . . . . 0 . . . . . .
76	1	6 0 2 . 0 . 0 0 . 0 0 0 0 0 0 0 0 0 0 0 1 . . 2 . . . 1 0 . 0
77	2	1 . . 0 . . 0 . . 0 . . 0 . . 0 . . . . . 0 . 0 . . . 0 . . .
78	4	1 0 1 0 . . 0 0 0 0 . 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . . 0 . .
79	4	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
80	3	3 . . . . 0 . . . 0 . . . . 0 . . . . 0 .
81	1	4 . . . 1 . . . 0 . . . . 0 . . . . 0 . . 1 . . . . . .
82	1	1 0 0 . . . . 0 . . . . 0 . . . . .
83	1	2 0 . 0 0 0 0 . 0 . . 0 . 0 0 . . 0 . . 0 . . 0 . . 0 . . 0 .
84	4	3 0 0 0 0 0 0 0 0 0 0 0 . 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
85	3	1 0 0 0 0 0 0 0 0 0 0 0 0 0 . 0 . 0 0 . 0 0 0 . 0 0 0 0 0 0 0

Data set V (Continued)

Patient												Months											
ID	31									40										50			53
Placebo group																							
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
26	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
27	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
28	0	.	0	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
29	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
30	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
31	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
32	0	.	.	.	.	0	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.
33	.	0	.	1	.	.	0	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.
34	.	.	.	.	0	0	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.
35	.	.	.	0	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.
36	.	0	.	.	0	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.
37	.	.	0	.	.	0	.	.	0	.	0	.	.	0	.	.	.	.	.	.	.	.	.
38	.	0	.	.	0	.	.	.	0	.	.	0	.	.	0	.	.	.	.	.	.	.	.
39	0	.	.	0	.	.	0	.	0	.	.	.	.	.	0	.	.	0	.	.	.	.	.
40	.	.	.	.	0	.	.	.	.	.	0	.	.	.	0	.	.	.	0	.	.	.	.
41	.	.	.	.	1	.	.	.	.	.	.	.	.	0	.	.	.	0	.	.	0	.	.
42	.	.	0	.	0	.	0	.	.	.	0	.	.	0	.	.	0	.	.	.	.	.	0

Data set V (Continued)

Patient	Months																						
ID	31									40										50			53
43	0	.	.	.	0	.	.	.	.	.	0	.	.	.	.	3	.	.	.	.	2	.	1
44	.	.	.	.	0	.	.	.	.	.	0	.	0	.	.	.	0	.	.	.	.	.	.
45	.	.	.	4	.	.	0	.	1	.	.	.	1	.	.	0	.	.	1	.	.	1	.
46	.	.	0	.	.	0	.	.	.	.	9	.	.	0	0	0	.	.	0	.	.	.	0
47	.	.	1	.	.	.	0	.	.	.	0	.	.	.	0	.	.	.	1	.	.	0	.
Thiotepa group																							
48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
49	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
51	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
52	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
53	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
54	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
55	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
56	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
57	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
58	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
59	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
60	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
61	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
62	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
63	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
64	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
65	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
66	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
67	.	0	.	.	3	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
68	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
69	.	.	8	.	0	9	8	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
70	.	3	.	0	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
71	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.
72	0	0	0	0	.	0	0	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.
73	0	0	0	0	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.
74	.	0	0	.	.	0	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.
75	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.
76	.	.	.	.	.	.	.	8	.	.	0	.	.	0	.	.	.	.	.	.	.	.	.
77	.	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.
78	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.	.	.	.	.	.	.
79	.	0	0	0	0	0	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.
80	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	0	.	.	.	.
81	0	.	0	.	.	0	.	.	.	.	.	.	0	.	.	.	1	.	.	0	.	.	.
82	.	0	.	.	.	.	.	0	.	.	.	.	.	0	.	.	0	.	.	0	.	.	.
83	.	0	.	.	.	.	.	2	.	.	0	.	.	.	0	.	.	0	.	.	0	.	.
84	0	0	0	0	.	0	.	.	0	.	.	0	.	.	0	.	.	0	.	.	.	.	.
85	0	0	0	0	0	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.

References

Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. *Spring Lect. Notes Statist.* **2**, 1-25. Mathematical Statistics and Probability Theory. W. Klonecki, A. Kozek, and J. Rosiński, editors.

Alioum, A. and Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, **52**, 512-524.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer-Verlag: New York.

Andersen, P. K. and Gill, R. D. (1982). Cox Regression Model for Counting Processes: a Large Sample Study. *The Annals of Statistics*, **10**, 1100-1120.

Andersen, P. K. and Ronn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometrics*, **51**, 323-329.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, **26**, 641-647.

Babineau, D. (2005). Goodness of fit for lifetime data models when responses are interval-censored. *Ph.D. Thesis*, University of Waterloo, Waterloo, Ontario, Canada.

Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association*, **85**, 1002-1008.

Bacchetti, P. and Quale, C. (2002). Generalized additive models with interval-censored data and time-varying covariates: application to human ummunodeficiency virus infection in hemophiliacs. *Biometrics*, **58**, 443-447.

Bagdonavicius, V. and Nikulin, M. S. (2001). Accelerated life models: modeling and statistical analysis. Chapman & Hall: London; New York

Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scandinavian Journal of Statistics*, **32**, 405-424.

Banerjee, S. and Carlin, B. P. (2004). Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, **60**, 268-275.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under order restrictions*. New York: Wiley.

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, **75**, 65-74.

Bean, S. J. and Tsokos, C. P. (1980). Developments in non-parametric density estimation. *Int. Statist. Rev.*, **48**, 267-287.

Bebchuk, J. D. and Betensky, R. A. (2000). Multiple imputation for simple estimation of the hazard function based on interval censored data. *Statistics in Medicine*, **19**, 405-419.

Becker, N. G. and Melbye, M. (1991). Use of a log-linear model to computer the empirical survival curve from interval-censored data, with application to data on test for HIV positivity. *Australian Journal of Statistics*, **33**, 125-133.

Betensky, R. A. (2000). On nonidentifiability and noninformative censoring for current status data. *Biometrika*, **87**, 218-221.

Betensky, R. A. and Finkelstein, D. M. (1999a). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine*, **18**, 3101-3109.

Betensky, R. A. and Finkelstein D. M. (1999b). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, **18**, 3089-3100.

Betensky, R. A. and Finkelstein, D. M. (2002). Testing for dependence between failure time and visit compliance with interval-censored data. *Biometrics*, **58**, 58-63.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics*, **55**, 238-245.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, **21**, 263-275.

Betensky, R. A., Rabinowitz, D. and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika*, **88**, 703-711.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.

Boardman, T. J. (1973). Estimation in compound exponential failure models - when the data are grouped. *Technometrics*, **12**, 271-277.

Bogaerts, K., Leroy, R., Lesaffre, E. and Declerck, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in Medicine*, **21**, 3775-3787.

Bogaerts, K. and Lesaffre, E. (2004). A new, fast algorithm to find the regions of possible support for bivariate interval-censored data. *Journal of Computational Graphical Statistics*, **13**, 330-340.

Böhning, D., Schlattmann, P. and Dietz, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika*, **83**, 462-466.

Braun, J, Duchesne, T. and Stafford, J. E. (2005). Local likelihood density estimation for interval censored data. *The Canadian Journal of Statistics*, **33**, in press.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38-44.

Breslow, N. E. and Day, N. E. (1987). *Statistical methods in cancer research*, **2**, *The design and analysis cohort studies*. Lyon: IARC.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society, B*, **43**, 41-45.

Burridge, J. (1982). Some unimodality properties of likelihoods derived from grouped data. *Biometrika*, **69**, 145-151.

Byar, D. P. (1980). The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparison of placebo, pyridoxine, and topical thiotepa. In *Bladder Tumors and Other Topics in Urological Oncology*, eds. Pavone-Macaluso, M., Smith, P. H. and Edsmyn, F., New York: Plenum, 363-370.

Byar, D. P., Blackard, C. and The Veterans Administration Cooperative Urological Research Group (1977). Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage I bladder cancer. *Urology*, **10**, 556-561.

Cai, J. and Kim, J. (2003). Nonparametric quantile estimation with correlated failure time data. *Lifetime Data Analysis*, **9**, 357-371.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, **82**, 151-164.

Cai, J. and Schaubel, D. E. (2004). Analysis of recurrent event data. *Handbook of Statistics* **23**, 603-623.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, **59**, 570-579.

Cai, T. and Cheng, S. (2004). Semiparametric regression analysis for doubly censored data. *Biometrika*, **91**, 277-290.

Calle, M. L. and Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference*, **98**, 73-87.

Calle, M. L. and Gómez, G. (2005). A semiparametric hierarchical method for a regression model with an interval-censored covariate. *Australian & New Zealand Journal of Statistics*, **47**, 351-364.

Carstensen, B. (1996). Regression models for interval censored survival data: application to HIV infection in Danish homosexual men. *Statistics in Medicine*, **15**, 2177-2189.

Chang, S. and Wang, M-C. (1999). Conditional regression analysis for recurrence time data. *Journal of the American Statistical Association*, **94**, 1221-1230.

Chen, B. E. and Cook, R. J. (2003). Regression modeling with recurrent events and time-dependent interval-censored marker data. *Lifetime Data Analysis*, **9**, 275-291.

Chen, B. E., Cook, R. J., Lawless, J. F. and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine*, **24**, 671-691.

Chen, H. (2001). Weighted Semiparametric Likelihood Method for Fitting a Proportional Odds Regression Model to Data From the Case-Cohort Design. *Journal of the American Statistical Association*, **96**, 1446-1457.

Chen, K., Jin, Z. and Ying. Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659-668.

Chen, K. and Zhou, M. (2004). Nonparametric hypothesis testing and confidence intervals with doubly censored data. *Lifetime Data Analysis*, **9**, 71-91.

Chen, Y. Q., Wang, M-C. and Huang, Y. (2004). Semiparametric regression analysis on longitudinal pattern of recurrent gap times. *Biostatistics*, **5**, 277-290.

Cheng, S. C. and Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika*, **87**, 89-97.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 835-845.

Cheng, S. C., Wei, L. J. and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association*, **92**, 227-235.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141-151.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

Collett, W. E. (1994). *Modelling survival data in medical research.* London: Chapman and Hall.

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *Statistics and Operations Research Transactions*, **27**, 1-11.

Cook, R. J. and Lawless, J. F. (1996). Interim monitoring of longitudinal comparative studies with recurrent event responses. *Biometrics*, **52**, 1311-1323.

Cook, R. J. and Lawless, J. F. (2006). *The analysis of recurrent event data.* Springer-Verlag: New York.

Cook, R. J., Lawless, J. F. and Nadeau, J. C. (1996). Robust tests for treatment comparisons based on recurrent event responses. *Biometrics*, **52**, 557-571.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data.*, Chapman & Hall: London.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248-275.

Datta, S., Satten, G. A. and Williamson, J. M. (2000). Consistency and asymptotic normality of estimators in a proportional hazards model with interval censoring and left truncation. *Annals of the Institute of Statistical Mathematics*, **52**, 160-172.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1-12.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Dewanji, A. and Kalbfleisch, J. D. (1986). Nonparametric methods for survival/sacrifice experiments. *Biometrics* 42, 325-341.

Diamond, I. D. and McDonald, J. W. (1991). The analysis of current status data. *Demographic Applications of Event History Analysis*, eds. Trussel, J., Hankinson, R. and Tilton, J. Oxford University Press: Oxford, U.K.

Diamond, I. D., McDonald, J. W. and Shah, I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, **23**, 607-620.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *The analysis of longitudinal data.* Oxford University Press Inc., New York.

Ding, A. A. and Wang, W. (2004). Testing independence for bivariate current status data. *Journal of the American Statistical Association*, **99**, 145-155.

Dinse, G. E. (1994). A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments. *Statistics in Medicine*, **13**, 689-708.

Dinse, G. E. and Lagakos, S. W. (1983). Regression analysis of tumor prevalence data. *Applied Statistics*, **32**, 236-248.

DiRienzo, A. G. (2003). Nonparametric comparison of two survival-time distributions in the presence of dependent censoring. *Biometrics*, **59**, 497-504.

Dorey, F. J. , Little, Roderick J. A. , and Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, **12**, 1589-1603.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, **22**, 1763-1786.

Dunson, D. B. and Dinse, G.. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics*, **58**, 79-88.

Dunson, D. B., Holloman, C., Calder, C. and Gunn, L. H. (2004). Bayesian modeling of multiple lesion onset and growth from interval-censored data. *Biometrics*, **60**, 676-683.

Efron, B. (1967). The two sample problem with censored data. In *Proc. 5th Berkeley Symp. on Math. Statist. Prob.*. Berkeley: University of California Press, 831-853.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. Second Edition, Marcel Dekker, Inc., New York.

Fang, H. and Sun J. (2001). Consistency of nonparametric maximum likelihood estimation of a distribution function based on doubly interval-censored failure time data. *Statistics and Probability Letters*, **55**, 311-318. *Biometrics*, **51**, 502-511.

Fang, H., Sun, J. and Lee, M-L T. (2002). Nonparametric survival comparison for interval-censored continuous data, *Statistica Sinica*, **12**, 1073-1083.

Farrington, C. P. (1996). Interval censored survival data: A generalized linear modelling approach. *Statistics in Medicine*, **15**, 283-292.

Farrington, C. P. (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics*, **56**, 473-482.

Fay, M. P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics*, **52**, 811-822.

Fay, M. P. (1999a). Comparing several score tests for interval censored data. *Statistics in Medicine*, **18**, 273-285.

Fay, M. P. (1999b). Splus functions for nonparametric estimation and treatment comparison. http://lib.stat.cmu.edu/S/interval.tar.gz.

Fay, M. P. and Shih, J. H. (1998). Permutation tests using estimated distribution functions. *Journal of the American Statistical Association*, **93**, 387-396.

Feller, W. (1971). *An introduction to probability theory*. 2nd ed., Vol. II. New York: Wiley.

Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. *The Annals of Statistics*, **1**, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615-629.

Fine, J. P. and Jiang, H. (2000). On association in a copula with time transformations. *Biometrika*, **87**, 559-571.

Fine, J. P. Ying, Z. and Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika*, **85**, 980-986.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845-854.

Finkelstein, D. M., Goggins, W. B. and Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring. *Biometrics*, **582**, 298-304.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, **41**, 933-945.

Finkelstein, D. M. and Wolfe, R. A. (1986). Isotonic Regression for interval censored survival data using an E-M algorithm. *Communications in Statistics: Theory and Methods*, **15**, 2493-2505.

Fleming, T. R. and Harrington, D. P. (1991). *Counting process and survival analysis*. John Wiley: New York.

Freireich, E. O. et al. (1963). The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia., *Blood*, **21**, 699-716.

Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society, Series B*, **54**, 853-866.

Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society, Series B*, **56**, 71-74.

Frydman, H. (1995a). Nonparametric estimation of a Markov "illness-death" process from interval-censored observations, with application to diabetes survival data. *Biometrika*, **82**, 773-789.

Frydman, H. (1995b). Semiparametric estimation in a three-state duration-dependent Markov model from interval-censored observations with application to AIDS data. *Biometrics*, **51**, 502-511.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analyses of event rates. *Technometrics*, **29**, 1-15.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203-223.

Genest, C. and Rivest, L. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, **88**, 1034-1043.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, **81**, 618-623.

Gentleman, R. and Vandal, A. C. (2001). Computational algorithms for censored data problems using intersection graphs. *Journal of Computational Graphical Statistics*, **10**, 403-421.

Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data. *The Canadian Journal of Statistics*, **30**, 557-571.

Geskus, R. and Groeneboom, P. (1996). Asymptotically optimal estimation of smooth functionals for interval censoring, part 1. *Statistica Neerlandica*, **50**, 69-88.

Geskus, R. and Groeneboom, P. (1997). Asymptotically optimal estimation of smooth functionals for interval censoring, part 1. *Statistica Neerlandica*, **50**, 201-219.

Geskus, R. and Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *The Annals of Statistics*, **27**, 627-674.

Ghosh, D. (2001). Efficiency considerations in the additive hazards model with current status data. *Statistica Neerlandica*, **55**, 367-376.

Ghosh, D. (2003). Goodness-of-fit methods for additive-risk models in tumorgenicity experiments. *Biometrics*, **59**, 721-726.

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at random: characterizations, conjectures, counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. Lin, D. and Fleming, T. Springer-Verlag, New York, 255-294.

Goedert, J., Kessler, C. Adedort, L. and et al. (1989). A prospective-study of human immunodeficiency virus type-1 infection and the development of AIDS in subjects with hemophilia. *New England Journal of Medicine*, **321**, 1141-1148.

Goggins, W. B. , and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics*, **56**, 940-943.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A. and Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval censored data under the Cox proportional hazards model. *Biometrics*, **54**, 1498-1507.

Goggins, W. B., Finkelstein, D. M. and Zaslavsky, A. M. (1999a). Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine*, **18**, 2737-2747.

Goggins, W. B., Finkelstein, D. M. and Zaslavsky, A. M. (1999b). Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval-censored. *Biometrics*, **55**, 445-451.

Gómez, G. and Calle, M. L. (1999). Nonparametric estimation with doubly censored data. *Journal of Applied Statistics*, **26**, 45-58.

Gómez, G., Calle, M. L., Egea, J. M. and Muga, R. (2000). Estimation of the risk of HIV infection as a function of the length of intravenous drug use: A nonparametric Bayesian approach. *Statistics in Medicine*, **19**, 2641-2656.

Gómez, G., Calle, M. L. and Oller, R. (2004). Frequentist and Bayesian approaches for interval-censored data. *Statistics Papers*, **2**, 139-173.

Gómez, G., Espinal and Lagakos, S. W. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, **22**, 409-425.

Gómez, G. and Lagakos, S. W. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*, **50**, 204-212.

Goodall, R. L., Dunn, D. T. and Babiker, A. G. (2004). Interval-censored survival time data: confidence intervals for the nonparametric survivor function. *Statistics in Medicine*, **23**, 1131-1145.

Greenwood M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, **33**, 1-26.

Groeneboom, P. (1995). Nonparametric estimators for interval censoring problems. *Analysis of Censored Data (Pune, 1994/1995)*, eds. H. L. Koul and J. V. Deshoande, IMS Lecture Notes, Monograph Series **27**, 105-128.

Groeneboom, P. (1996). Lectures on inverse problems. In *Lecture Notes in Mathematics*, 1648, Springer-Verlag: Berlin.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* DMV Seminar, Band 19, Birkhauser, New York.

Groeneboom, P. and Wellner, J. A. (2001). Computing Chernoff's distribution. *Journal of Computational & Graphical Statistics*, **10**, 388-400.

Groenewald, P. C. N. and Mokgtlhe, L. (2004). Bayesian analysis of bivariate interval censored survival data. *South African Statistical Journal*, **38**, 25-45.

Gulick, R. M., Hu, X. J., Fiscus, S. A., Fletcher, C. V., Haubrish, R., Cheng, H., Acosta, E., Lagakos, S. W., Swanstrom, R., Freimuch, W., Snyder, S., Mills, C., Fischl, M., Pettinelli, C. and Katzenstein, D (2000). Randomized study of saquinavir with ritonavir or nelfinavir together with delavirdine, adefovir, or both in human immunodeficiency virus-infected-adults with virologic failure on indinavir: AIDS Clinical Trials Group study 359. *Journal of Infectious Diseases*, **182**, 1375-1384.

Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, **50**, 632-639.

Grummer-Strawn, L. M. (1993). Regression analysis of current status data: An application to breast-feeding. *Journal of the American Statistical Association*, **88**, 758-765.

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data. *Journal of Computational and Graphical Statistics*, **13**, 341-361.

Hazelrig, J. B., Turner, M. E. and Blackstone, E. H. (1982). Parametric survival analysis combining longitudinal and cross-sectional-censored and interval-censored data with cooncomitant information. *Biometrics*, **38**, 1-15.

He, W. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics*, **59**, 837-848.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, **19**, 2244-2253.

Hinde, J. (1982). Compound Poisson regression models. In *GLIM 82: Proceedings of the International Conference in Generalized Linear Models*, R. Gilchrist (ed), Berlin: Springer-Verlag, 109-121.

Hjort, N. L. (1990). Nonparametric Bayesian estimators based on beta processes in models of life history data. *The Annals of Statistics*, **18**, 1259-1294.

Hoel, D. G. and Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute*, **49**, 361-372.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, **73**, 671-678.

Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer-Verlag: New York.

Hsu, L. and Prentice, R. L. (1996). On assessing the strength of dependency between failure time variates. *Biometrika*, **83**, 491-506.

Hu, X. J., Sun, J. and Wei, L. J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics*, **30**, 25-43.

Huang, C. Y. and Wang, M-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *Journal of the American Statistical Association*, **99**, 1153-1165.

Huang, J. (1995). Maximum likelihood estimation for proportional odds regression model with current status data. *Analysis of Censored Data*, IMS Lecture Notes - Monograph Series 27, 129-146.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, **24**, 540-568.

Huang, J. (1999a). Efficient estimation of the partly linear additive Cox model. *The Annals of Statistics*, **27**, 1536-1563

Huang, J. (1999b). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, **9**, 501-519.

Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, **92**, 960-967.

Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statistics Neerlandica*, **49**, 153-163.

Huang, J. and Wellner, J. A. (1996). Regression models with interval censoring. *Probability Theory and Mathematical Statistics (St. Petersburg, 1993)*, 269-296.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. Lin, D. and Fleming, T. Springer-Verlag, New York, 123-169.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, **58**, 510-520.

Huang, Y. and Chen, Y. Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime Data Analysis*, **9**, 293-303.

Huber-Carol, C. and Vonta, I. (2004). Frailty models for arbitrarily censored and truncated data. *Lifetime Data Analysis*, **10**, 369-388.

Hudgens, M. G. (2005). On nonparametric maximum likelihood estimation with interval censoring and left truncation. *Journal of the Royal Statistical Society, Series B*, **67**, 573-587.

Hudgens, M. G., Satten, G. A. and Longini, I. M. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics*, **57**, 74-80.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian survival analysis.* Springer-Verlag: New York.

Ii, Y., Kikuchi, R., and Matsuoka, K. (1987). Two-dimensional (time and Multiplicity) statistical analysis of multiple tumors. *Mathematical Bioscience*, **84**, 1-21.

Ishwaran, H. and James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *Journal of the American Statistical Association*, **99**, 175-190.

Jewell, N. P. (1994). Non-parametric estimation and doubly-censored data: general ideas and applications to AIDS. *Statistics in Medicine*, **13**, 2081-2095.

Jewell, N. P., Malani, H. M. and Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association*, **89**, 7-18.

Jewell, N. P. and Shiboski, S. C. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics*, **46**, 1133-1150.

Jewell, N. P. and van der Laan, M. J. (1996). Generalizations of current status data with applications. *Lifetime Data: models in Reliability and Survival Analysis*, Kluwer Acad. Publ., Dordrecht, 141-148.

Jewell, N. P. and van der Laan, M. J. (1997). Singly and doubly censored current status data with extensions to multi-state counting processes. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. Lin, D. and Fleming, T. Springer-Verlag, New York, 171-184.

Jewell, N. P. and van der Laan, M. J. (2004a). Current status data: review, recent developments and open problems. *Advances in Survival Analysis*, Elsevier, Amsterdam, 625-642.

Jewell, N. P. and van der Laan, M. J. (2004b). Case-control current status data. *Biometrika*, **91**, 529-541.

Jewell, N. P., van der Laan, M. J. and Lei, X. (2005). Bivariate current status data with univariate monitoring times. *Biometrika*, **92**, 847-862.

Jewell, N. P., van der Laan, M. J. and Hennemean, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, **90**, 183-197.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. *Biometrics*, **55**, 887-890.

Joly, P., Commenges, D. and Letenneeur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence dementia. *Biometrics*, **54**, 185-194.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, **7**, 310-321.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**, 214-221.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863-871.

Kalbfleisch, J. D. , Lawless, J. F. and Robinson, J. A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, **33**, 273-285.

Kalbfleisch, J. D. and MacKay, R. J. (1979). On constant-sum models for censored survival data. *Biometrika*, **66**, 87-90.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267-278.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. Second edition, John Wiley: New York.

Kang, S. and Koehler, K. J. (1998). Maximum likelihood estimation of bivariate survival probabilities and its existence and uniqueness. *Communication in Statistics - Theory and Methods*, **27**, 1961-1977.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A*, **154**, 371-412.

Keiding, N., Begtrup, K. Scheike, T. H. and Hasibeder, G. (1996). Estimation from current status data in continuous time. *Lifetime Data Analysis*, **2**, 119-129.

Kim, J. and Lee, S. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika*, **85**, 593-603.

Kim, J. S. (2003a). Efficient estimation for the proportional hazards model with left-truncation and case-I interval-censored data. *Statistica Sinica*, **13**, 519-537.

Kim, J. S. (2003b). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society, Series B*, **65**, 489-502.

Kim, M. Y., De Gruttola, V. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, **49**, 13-22.

Kim, M. Y. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, **21**, 3715-3726.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis*, Springer-Verlag: New York.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, **61**, 387-421.

Kong, L., Cai, J. and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*, **91**, 305-319.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics*, **53**, 1485-1494.

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, **1**, 301-328.

Kroner, B. L., Rosenberg, P. S., Aledort, L. M., Alvord, W. G. and Goedert, J. J. (1994). HIV-1 infection incidence among persons with hemophilia in the United States and Western Europe, 1978-1990. *Journal of Acquired Immune Deficiency Syndromes*, **7**, 279-286.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association*, **95**, 238-248.

Lagakos, S. W. (1980). The graphical evaluation of explanatory variables in proportional hazards regression. *Biometrika*, **68**, 93-98.

Lagakos, S. W. and Louis, T. A. (1988). Use of tumor lethality to interpret tumorgenicity experiments lacking cause-of-death data. *Applied Statistics*, **37**, 169-179.

Lagakos, S. W. and Williams, J. S. (1978). Models for censored survival analysis: a cone class of variable-sum models. *Biometrika*, **65**, 181-189.

Lam, K. F. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, **92**, 573-586.

Langohr, K., Gómez, G. and Muga, R. (2004). A parametric survival model with an interval-censored covariate. *Statistics in Medicine*, **23**, 3159-3175.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data.* John Wiley: New York.

Lawless, J. F. (2004). A note on interval-censored lifetime data and the constant-sum condition of Oller, Gómez & Calle. *Canadian Journal of Statistics*, **32**, 327-331.

Lawless, J. F. and Nadeau, J. C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events. *Technometrics*, **37**, 158-168.

Lawless, J. F., Wigg. M. B., Tuli, S., Drake, J. and Lamberti-Pasculli, M. (2001). Analysis of repeated failures or durations, with application to shunt failures for patients with paediatric hydrocephalus. *Journal of the Royal Statistical Society, Series C*, **50**, 449-465.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics*, **26**, 549-565.

Leung, M. K. and Elashoff, R. M. (1996). Generalized linear mixed-effects models with a finite-support random-effects distribution: a maximum-penalized-likelihood approach. *Biometrical Journal*, **38**, 135-151.

Li, L. (2003). Linear regression analysis with observations subject to interval censoring. *Development of Modern Statistics and Related Topics*, World Sci. Publishing, River Edge, NJ, 236-245.

Li, L. and Pu. Z. (1999). Regression models with arbitrarily interval-censored observations. *Communications in Statistics. Theory and Methods*, **28**, 1547-1563.

Li, L. and Pu. Z. (2003). Rank estimation of log-linear regression with interval-censored data. *Lifetime Data Analysis*, **9**, 57-70.

Li, L., Watkins, T. and Yu, Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, **24**, 531-542.

Li, Q. H. and Lagakos, S. W. (2004). Comparisons of test statistics arising from marginal analyses of multivariate survival data. *Lifetime Data Analysis*, **10**, 389-405.

Lim, H. J. and Sun, J. (2003). Nonparametric tests for interval-censored failure time data. *Biometrical Journal*, **45**, 263-276.

Lim, H. J., Sun J. and Matthews D. E. (2002). Maximum likelihood estimation of a survival function with a change-point for truncated and interval-censored data. *Statistics in Medicine*, **21**, 743-752.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**, 2233-2247.

Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, **85**, 289-298.

Lin, D. Y., Robins, J. M. and Wei, L . J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*, **83**, 381-393.

Lin, D. Y. , Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, **80**, 557-572.

Lin, D. Y. , Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B*, **62**, 711-730.

Lin, D. Y. , Wei, L. J. and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, **85**, 605-618.

Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, **80**, 573-581.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *The Annals of Statistics*, **23**, 1712-1734.

Lin, D. Y. and Ying, Z. (1997). Additive hazards regression models for survival data. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, eds. Lin, D. and Fleming, T. Springer-Verlag, New York, 185-198.

Lin, H. and Scharfstein, D. O. (2004). Analysis of longitudinal data with irregular outcome-dependent follow-up. *Journal of Royal Statistical Society, Series B*, **66**, 791-813.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: methods for interval-censored data. *Statistics in Medicine*, **17**, 219-238.

Lindsey, J. K. (1998). A study of interval censoring in parametric regression models. *Lifetime Data Analysis*, **4**, 329-354.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112-1121.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* Wiley, New York.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, **91**, 331-343.

Maathuis, M. H. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval-censored data. *Journal of Computational and Graphical Statistics*, **14**, 352-362.

Mantel, N. (1967). Ranking procedures for arbitrarily restricted observations. *Biometrics*, **23**, 65-78.

Marshall, M. L. (1974). Fitting the two-term mixed exponential and two-parameter lognormal distributions to grouped and censored data. *Applied Statistics*, **23**, 313-322.

Martinussen, T. and Scheike, T. H. (2002a). A flexible additive multiplicative hazard model. *Biometrika*, **89**, 283-298.

Martinussen, T. and Scheike, T. H. (2002b). Efficient estimation in additive hazards regression with current status data. *Biometrika*, **89**, 649-658.

Murphy, S. A., Rossini, A. J. and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, **92**, 968-976.

Nelson, W. B. (2003). *Recurrent events data analysis for product repairs, disease recurrences, and other applications.* ASA-SIAM Series on Statistics and Applied Probability, **10**.

Ng, M. P. (2002). A modification of Peto's nonparametric estimation of survival curves for interval-censored data. *Biometrics*, **58**, 439-442.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, **44**, 414-422.

Odell, P. M. , Anderson, K. M. , and D'Agostino, R. B. (1992) Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, **48**, 951-959.

Oller, R., Gómez, G. and Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics*, **32**, 315-326.

Pan, W. (1999). A comparison of some two-sample tests with interval-censored data. *Journal of Nonparametric Statistics*, **12**, 133-146.

Pan, W. (2000a). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, **56**, 199-203.

Pan, W. (2000b). A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, **19**, 1-11.

Pan, W. (2000c). Smooth estimation of the survival function for interval censored data. *Statistics in Medicine*, **19**, 2611-2624.

Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics*, **57**, 1245-1250.

Pan, W. and Chappell, R. (1998a). Estimating survival curves with left truncated and interval censored data via the EMS algorithm. *Communications in Statistics. Theory and Methods*, **27**, 777-793.

Pan, W. and Chappell, R. (1998b). Estimating survival curves with left truncated and interval-censored data under monotone hazards. *Biometrics* **54**, 1053-1060.

Pan, W. and Chappell, R. (1999). A note on inconsistency of NPMLE of the distribution function from left truncated and case I interval censored data. *Lifetime Data Analysis*, **5**, 281-291.

Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left truncated and interval censored data. *Biometrics* **58**, 64-70.

Pan, W., Chappell, R. and Kosorok, M. R. (1998). On consistency of the monotone MLE of survival for left truncated and interval-censored data. *Statistics and Probability Letters*, **38**, 49-57.

Park, Y. and Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, **90**, 717-723.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, **88**, 811-820.

Pepe, M. S. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497-507.

Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86-91.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A*, **135**, 185-207.

Petroni, G. R. and Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics*, **50**, 77-87.

Pierce, D. A., Stewart, W. H. and Kopecky, K. J. (1979). Distribution free regression analysis of grouped survival data. *Biometrics*, **35**, 785-793.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, **79**, 495-512.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, **34**, 57-67.

Prentice, R. L. and Kalbfleisch, J. D. (2003). Aspects of the analysis of multivariate failure time data. *Statistics and Operations Research Transactions*, **27**, 65-78.

Rabinowitz, D., Betensky, R. A. and Tsiatis, A. A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics*, **56**, 511-518.

Rabinowitz, D. and Jewell, N. P. (1996). Regression with doubly censored current status data. *Journal of the Royal Statistical Society, Series B*, **58**, 541-550.

Rabinowitz, D., Tsiatis, A. A. and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, **82**, 501-513.

Rao, C. R. (1973). *Linear statistical inference and its applications.* 2nd ed. John Wiley & Sons, New York.

Ren, J. J. (2003). Goodness of fit tests with interval censored data. *Scandinavian Journal of Statistics*, **30**, 211-226.

Richman, D. D., Grimes, J. M. and Lagakos, S. W. (1990). Effect of stage of disease and drug use on zidovudine susceptibilities of isolates of human immunodeficiency virus. *Journal of AIDS*, **3**, 743-746.

Robertson, T. , Wright, F. T. and Dykstra, R. (1988). Order restricted statistical inference. John Wiley: New York.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, **56**, 779-788.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106-121.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, **51**, 874-887.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, **91**, 713-721.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, **82**, 805-820.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley: New York.

Rücker, G. and Messerer, D. (1988). Remission duration: An example of interval censored observations. *Statistics in Medicine*, **7**, 1139-1145.

Samuelson, S. O. and Kongerud, J. (1994). Interval censoring in longitudinal data of respiratory symptoms in aluminium potroom workers: a comparison of methods. *Statistics in Medicine*, **13**, 1771-1780.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval-censored data. *Biometrika*, **83**, 355-370.

Satten, G. A., Datta, S. and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association*, **93**, 318-327.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, **89**, 617-634.

Scheike, T. H. and Zhang, M. (2002). An additive-multiplicative Cox-Aalen regression model. *Scandinavian Journal of Statistics*, **29**, 75-88.

Schick, A. and Yu, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

Self, S. G. and Grossman, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*, **42**, 521-530.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*, **85**, 165-177.

Shen, X. (2000). Linear regression with current status data. *Journal of the American Statistical Association*, **95**, 842-852.

Shiboski, S. C. (1998). Generalized additive models for current status data. *Lifetime Data Analysis*, **4**, 29-50.

Shiboski, S. C. and Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, **87**, 360-372.

Shih, J. H. and Louis, T. A. (1995). Inference on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Shorack, G. R. (2000). *Probability for statisticians.* New York: Springer-Verlag.

Sinha, D. (1997). Time-discrete beta-process model for interval-censored survival data. *The Canadian Journal of Statistics*, **25**, 445-456.

Sinha, D., Chen, M.-H. and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics*, **55**, 585-590.

Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, 1195-1212.

Sinha, D. and Maiti, T. (2004). A Bayesian approach for the analysis of panel-count data with dependent termination. *Biometrics*, **60**, 34-40.

Song, S. (2004). Estimation with univariate "mixed case" interval censored data. *Statistica Sinica*, **14**, 269-282.

Staniswalls, J. G., Thall, P. F. and Salch, J. (1997). Semiparametric regression analysis for recurrent event interval counts. *Biometrics*, **53**, 1334-1353.

Sternberg, M. R. and Satten, G. A. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval-censoring and truncation. *Biometrics*, **55**, 514-522.

Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics*, **51**, 1096-1104.

Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine*, **15**, 1387-1395.

Sun, J. (1997a). Regression analysis of interval-censored failure time data. *Statistics in Medicine*, **16**, 497-504.

Sun, J. (1997b). Self-consistency estimation of distributions based on truncated and doubly censored data with applications to AIDS cohort studies. *Lifetime Data Analysis*, **3**, 305-313.

Sun, J. (1998). Interval censoring. *Encyclopedia of Biostatistics*, John Wiley, First Edition, 2090-2095.

Sun, J. (1999). A nonparametric test for current status data with unequal Censoring. *Journal of the Royal Statistical Society, Series B*, **61**, 243-250.

Sun, J. (2001a). Variance estimation of a survival function for interval-censored survival data. *Statistics in Medicine*, **20**, 1249-1257.

Sun, J. (2001b). Nonparametric test for doubly interval-censored failure time data. *Lifetime Data Analysis*, **7**, 363-375.

Sun, J. (2004). Statistical analysis of doubly interval-censored failure time data. Advances in survival analysis, *Handbook of Statistics*, **23**, 105-122.

Sun, J. (2005). Interval censoring. *Encyclopedia of Biostatistics*, John Wiley, Second Edition, 2603-2609.

Sun, J. and Fang, H. B. (2003). A nonparametric test for panel count data. *Biometrika*, **90**, 199-208.

Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, **88**, 1449-1454.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, **5**, 279-290.

Sun, J. and Kalbfleisch, J. D. (1996). Nonparametric tests of tumor prevalence data. *Biometrics*, **52**, 726-731.

Sun, J., Liao, Q. and Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics*, **55**, 909-914.

Sun, J., Lim, H. J. and Zhao, X. (2004). An independence test for doubly censored failure time data. *Biometrical Journal*, **46**, 503-511.

Sun, J. and Matthews, D. E. (1997). A random-effect regression model for medical follow-up studies. *The Canadian Journal of Statistics*, **25**, 101-111.

Sun, J., Park, D-H., Sun, L. and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, **100**, 882-889.

Sun, J. and Rai, S. N. (2001). Nonparametric tests for the comparison of point processes based on incomplete data. *Scandinavian Journal of Statistics*, **28**, 725-732.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *The Canadian Journal of Statistics*, **33**, 85-96.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B*, **62**, 293-302.

Sun, J., Zhao, Q. and Zhao, X. (2005). Generalized log rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics*, **32**, 49-57.

Sun, L., Kim Y. and Sun, J. (2004). Regression analysis of doubly censored failure time data using the additive hazards model. *Biometrics*, **60**, 637-643.

Sun, L., Park, D. and Sun, J. (2006). The additive hazards model for recurrent gap times. *Statistica Sinica*, in press.

Sun, L., Wang, L. and Sun, J. (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, in press.

Susarla, V. and van Ryzin, J. (1976), Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897-902.

Tang, M. X., Tsai, W. Y., Marder, K. and Mayeux, R. (1995). Linear rank tests for doubly censored data. *Statistics in Medicine*, **14**, 2555-2563.

Tanner, M. A. (1991). *Tools for statistical inference: observed data and data augmentation methods*. New York : Springer-Verlag.

Tanner, M. A. and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics*, **11**, 994-998.

Tanner, M. A. and Wong, W. H. (1987). The application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, **29**, 23-32.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics*, **44**, 197-209.

Thall, P. F. (1989). Correction to: "Mixed Poisson likelihood regression models for longitudinal interval count data." *Biometrics*, **45**, 197-209.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, **83**, 339-347.

Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147-160.

Thompson, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics*, **33**, 463-470.

Tibishirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559-567.

Tikhov, M. S. (2004). Statistical estimation based on interval censored data. *Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life*, Birkhuser Boston, Boston, MA, 211-218.

Topp, R. and Gómez, G. (2004). Residual analysis in linear regression models with an interval-censored covariate. *Statistics in Medicine*, **23**, 3377-3391.

Tsiatis, A. A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Science*, **27**, 20-22.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, **18**, 354-372.

Tu, X. M. (1995). Nonparametric estimation of survival distributions with censored initiating time, and censored and truncated terminating time:

application to transfusion data for acquired immune deficiency syndrome. *The Annals of Statistics*, **44**, 3-16.

Turnbull, B. W. (1974). Nonparametric estimation of survivorship function with doubly censored data. *Journal of the American Statistical Association*, **69**, 169-173.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.

Turnbull, B. W. and Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, **34**, 367-375.

Vandal, A. C., Gentleman, R. and Liu, X. (2005). Constrained estimation and likelihood intervals for censored data. *The Canadian Journal of Statistics*, **33**, 71-83.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics*, **21**, 14-44.

van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics*, **24**, 596-627.

van der Laan, M. J. and Andrews, C. (2000). The nonparametric maximum likelihood estimator in a class of doubly censored current status data models with application to partner studies. *Biometrika*, **87**, 61-71.

van der Laan, M. J. Bickel, P. J. and Jewell, N. P, (1997). Singly and doubly censored current status data: estimation, asymptotics and regression. *Scandinavian Journal of Statistics*, **24**, 289-307.

van der Laan, M. J. and Hubbard, A. (1997). Estimation with interval censored data and covariates. *Lifetime Data Analysis*, **3**, 77-91.

van der Laan, M. J. and Jewell, N. P, (2001). The NPMLE for doubly censored current status data. *Scandinavian Journal of Statistics*, **28**, 537-547.

van der Laan, M. J. and Jewell, N. P, (2003). Current status and right-censored data structures when observing a marker at the censoring time. Dedicated to the memory of Herbert E. Robbins. *The Annals of Statistics*, **31**, 512-535.

van der Laan, M. J. and Robins, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, **93**, 693-701.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer: New York.

Vermunt, J. K. (1997). *Log-linear models for event histories*. Sage Publications Inc: Newbury Park, CA.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, Chapman & Hall, London.

Wang, M-C. and Chen Y. Q. (2000). Nonparametric and semiparametric trend analysis of stratified recurrence time data. *Biometrics*, **56**, 789-794.

Wang, M-C. , Qin, J. , and Chiang, C-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, **96**, 1057-1065.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, **87**, 879-893.

Wang, Y. and Taylor, M. G. (2001). Jointly modeling longitudinal and event time data with application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association*, **96**, 895-905.

Wang, Z. and Gardiner, J. C. (1996). A class of estimators of the survival function from interval-censored data. *The Annals of Statistics*, **24**, 647-658.

Wei, G. C. G. and Tanner, M. A. (1991). Application of multiple imputation to the analysis of censored regression data. *Biometrics*, **47**, 1297-1309.

Wei, L. J. , and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, **79**, 653-661.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065-1073.

Wei, L. J., Ying, Z. and Lin, D. Y. (1990). Linear regression analysis of censored survival based on rank tests. *Biometrika*, **77**, 145-151.

Wellner, J. A. (1995). Interval censoring case 2: alternative hypotheses. *Analysis of Censored Data (Pune, 1994/1995)*, eds. H. L. Koul and J. V. Deshoande, IMS Lecture Notes, Monograph Series **27**, 271-219.

Wellner, J. A. and Zhan, Y. (1997). A hybird algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, **92**, 945-959.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annual of Statistics*, **28**, 779-814.

Wellner, J. A., Zhang, Y. and Liu, H. (2004). A semiparametric regression model for panel count data: when do pseudo-likelihood estimators become badly inefficient? *Proceedings of the Second Seattle Symposium in Biostatistics*, Springer, New York, 143-174.

Williams, J. S. and Lagakos, S. W. (1977). Models for censored survival analysis: constant-sum and variable-sum models. *Biometrika*, **64**, 215-224.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-339.

Xue, H., Lam, K. F. and Li, G. (2004). Sieve maximum likelihood estimation for semiparametric regression models with current status data. *Journal of the American Statistical Association*, **99**, 346-356.

Xue, H., Lam, K. F., Ben, C. and De Wolf, F. (2006). Semiparametric accelerated failure time regression analysis with application to interval-censored HIV/AIDS data. *Statistics in Medicine*, in press.

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, **94**, 125-136.

Ying, Z. (1990). Nonlinear stochastic approximation procedures for L_p loss functions. *The Annals of Statistics*, **18**, 1817-1828.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, **53**, 1199-1211.

Yu, A. K. F., Kwan, K. Y. W., Chan, D. H. Y. and Fong, D. Y. T. (2001). Clinical features of 46 eyes with calcified hydrogel intraocular lenses. *Journal of Cataract and Refractive Surgery*, **27**, 1596-1606.

Yu, Q., Li, L. and Wong, G. Y. C. (1998). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhy, Series A*, **60**, 184-197.

Yu, Q., Li, L. and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, **27**, 35-44.

Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1998a). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statistics and Probability Letters*, **37**, 223-228.

Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1998b). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics*, **26**, 619-627.

Yu, Q. and Wong, G. Y. C. (2003). Semi-parametric MLE in simple linear regression analysis with interval-censored data. *Communications in Statistics. Simulation and Computation*, **32**, 147-163.

Yu, Q., Wong, G. Y. C. and He, Q. (2000). Estimation of a joint distribution function with multivariate interval-censored data when the nonparametric MLE is not unique. *Biometrical Journal*, **6**, 747-763.

Yu, Q., Wong, G. Y. C. and Li, L. (2001). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Annals of the Institute of Statistical Mathematics*, **53**, 469-486.

Zeng, D., Cai, J. and Shen, Y. (2006). Semiparametric additive risks model for interval-censored data. *Statistica Sinica*, **16**, 287-302.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*, **89**, 39-48.

Zhang, Y. and Jamshidian, M. (2003). The gamma-frailty Poisson model for the nonparametric estimation of panel count data. *Biometrics*, **59**, 1099-1106.

Zhang, Y., Liu, W. and Wu, H. (2003). A simple nonparametric two-sample test for the distribution function of event time with interval censored data. *Journal of Nonparametric Statistics*, **16**, 643-652.

Zhang, Y., Liu, W. and Zhan, Y. (2001). A nonparametric two-sample test of the failure function with interval censoring case 2. *Biometrika*, **88**, 677-686.

Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, **24**, 1399-1407.

Zhang, Z., Sun, L., Zhao, X. and Sun, J. (2005). Regression analysis of interval censored failure time data with linear transformation models. *The Canadian Journal of Statistics*, **33**, 61-70

Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine*, **23**, 1621-1629.

Zhao, Q. and Sun, J. (2006). Semiparametric and nonparametric analysis of recurrent events with observation gaps. *Computational Statistics and Data Analysis*, in press.

Zhao, X., Lim, H-J. and Sun, J. (2005). Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring. *Journal of Statistical Planning and Inference*, **129**, 145-157.

Zhou, M. (2004). Nonparametric Bayes estimation of survival functions for doubly/interval censored data. *Statistica Sinica*, **14**, 533-546.

Zhu, C. and Sun, J. (2006). Variance estimation of a survival function with doubly censored failure time data. *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis, and Survival Analysis*, eds. Balakrishnan, N., Auget, J.-L., Mesbah, M. and Molenberghs, G., Springer, 229-239.

Index